High Performance Biocomputation

MITRE

JASON

High Performance Biocomputation

Study Leader:
Dan Meiron

Contributors:

Henry Abarbanel
Michael Brenner
Curt Callan
William Dally
David Gifford
Russell Hemley
Terry Hwa
Gerald Joyce

Steve Koonin
Herb Levine
Nate Lewis
Darrell Long
Roy Schwitters
Christopher Stubbs
Peter Weinberger
Hugh Woodin

March 2005

JSR-04-300

Approved for public release; distribution unlimited

JASON
The MITRE Corporation
7515 Colshire Drive
McLean, Virginia 22102-7508
(703) 983-6997

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information estimated to average 1 hour per response, including the time for review instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget. Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE March 2005	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE High Performance Biocomputation	5. FUNDING NUMBERS 13059022-IN
6. AUTHOR(S) Dan Meiron et al.	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) The MITRE Corporation JASON Program Office 7515 Colshire Drive McLean, Virginia 22102	8. PERFORMING ORGANIZATION REPORT NUMBER JSR-04-300
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Department of Energy Washington, DC 20528	10. SPONSORING/MONITORING AGENCY REPORT NUMBER JSR-04-300

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release	12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This section summarizes the conclusions and recommendations of the 2004 JASON summer study commissioned by the Department of Energy (DOE) to explore the opportunities and challenges presented by applying advanced computational power and methodology to problems in the biological sciences. JASON was tasked to investigate the current suite of computationally intensive problems as well as potential future endeavors. JASON was also tasked to consider how advanced computational capability and capacity could best be brought to bear on bioscience problems and to explore how different computing approaches such as Grid computing, supercomputing, cluster computing or custom architectures might map onto interesting biological problems

14. SUBJECT TERMS	15. NUMBER OF PAGES
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT SAR

Contents

1 EXECUTIVE SUMMARY

This section summarizes the conclusions and recommendations of the 2004 JASON summer study commissioned by the Department of Energy (DOE) to explore the opportunities and challenges presented by applying advanced computational power and methodology to problems in the biological sciences. JASON was tasked to investigate the current suite of computationally intensive problems as well as potential future endeavors. JASON was also tasked to consider how advanced computational capability and capacity[1] could best be brought to bear on bioscience problems and to explore how different computing approaches such as Grid computing, supercomputing, cluster computing or custom architectures might map onto interesting biological problems.

The context for our study is the emergence of information science as an increasingly important component of modern biology. Major drivers for this include the enormous impact of the human genome initiative and further large-scale investments such as DOE's GTL initiative, the DOE Joint Genomics Institute, as well as the efforts of other federal agencies as exemplified by the BISTI initiative of NIH. It should be noted too that the biological community is making increasing use of computation at the Terascale level (implying computational rates and dataset sizes on the order of Teraflops and Petabytes, respectively) in support of both theoretical and experimental endeavors.

Our study confirms that computation is having an important impact at every level of the biological enterprise. It has facilitated investigation of computationally intensive tasks such as the study of molecular interactions that

[1] Our definition of capability and capacity follows that adopted in the 2003 JASON report "Requirements for ASCI"[36]. That report defines capability as the maximum processing power possible that can be applied to a single job. Capacity represents the total processing power available from all machines used to solve a particular problem.

affect protein folding, analysis of complex biological machines, determination of metabolic and regulatory networks, modeling of neuronal activity and ultimately multi-scale simulations of entire organisms. Computation has also had a key role in the analysis of the enormous volume of data arising from activities such as high-throughput sequencing, analysis of gene expression, high-resolution imaging and other data-intensive endeavors. Some of these research areas are highly advanced in their utilization of computational capability and capacity, while others will require similar capability and capacity in the future.

JASON was asked to focus on possible opportunities and challenges in the application of advanced computation to biology. Our findings in this study are as follows:

Role of computation: Computation plays an increasingly important role in modern biology at all scales. High-performance computation is critical to progress in molecular biology and biochemistry. Combinatorial algorithms play a key role in the study of evolutionary dynamics. Database technology is critical to progress in bioinformatics and is particularly important to the future exchange of data among researchers. Finally, software frameworks such as BioSpice are important tools in the exchange of simulation models among research groups.

Requirements for capability: Capability is presently not a key limiting factor for any of the areas that were studied. In areas of molecular biology and biochemistry, which are inherently computationally intensive, it is not apparent that substantial investment will accomplish much more than an incremental improvement in our ability to simulate systems of biological relevance given the current state of algorithms. Other areas, such as systems biology will eventually be able to utilize capability computing, but the key issue there is out lack of understanding of more fundamental aspects, such as the details of cellular signaling processes.

Requirements for capacity: Our study did reveal a clear need for additional capacity. Many of the applications reviewed in this study (such as image analysis, genome sequencing, etc.) utilize algorithms that are essentially "embarrassingly parallel" algorithms and would profit simply from the increased throughput that could be provided by commodity cluster architecture as well as possible further developments in Grid technology.

Role of grand challenges: In order to elucidate possible applications that would particularly benefit from deployment of enhanced computational capability or capacity, JASON applied the notion of "grand challenges" as an organizing principle to determine the potential benefit of significant investment in either capability or capacity as applied to a given problem. JASON criteria for such grand challenges are as follows:

- they must be science driven;
- they must focus on a difficult but ultimately achievable goal;
- there must exist promising ideas on how to surmount existing limits;
- one must know when the stated goal has been achieved;
- the problem should be solvable in a time scale of roughly one decade;
- the successful solution must leave a clear legacy and change the field in a significant way.

These challenges are meant to focus a field on a very difficult but imaginably achievable medium-term goal. Some examples are discussed below in this summary as well as in the body of the report. It is plausible (but not assured) that there exist suitable grand challenge problems (as defined above) that will have significant impact on biology and which require high performance capability computing.

Future challenges: For many of the areas examined in this study, significant research challenges must be overcome in order to maximize the potential of high-performance computation. Such challenges include overcoming the complexity barriers in current biological modelling algorithms and understanding the detailed dynamics of components of cellular signaling networks.

JASON recommends that DOE consider four general areas in its evaluation of potential future investment in high performance bio-computation:

1. Consider the use of grand challenge problems, as defined above, to make the case for present and future investment in high performance computing capability. While some illustrative examples have been considered in this report, such challenges should be formulated through direct engagement with (and prioritization by) the bioscience community in areas such as (but not limited to) molecular biology and biochemistry, computational genomics and proteomics, computational neural systems, and systems or synthetic biology. Such grand challenge problems can also be used as vehicles to guide investment in focused algorithmic and architectural research, both of which are essential to achievement of grand challenge problems.
2. Investigate further investment in capacity computing. As stated above, a number of critical areas can benefit immediately from investments in capacity computing, as exemplified by today's cluster technology.
3. Investigate investment in development of a data federation infrastructure. Many of the "information intensive" endeavors reviewed here can be aided through the development and curation of datasets utilizing community adopted data standards. Such applications are ideally suited for Grid computing.
4. Most importantly, while it is not apparent that capability computing is, at present, a limiting factor for biology, we do not view this situation as static and, for this reason, it is important that the situation

be revisited in approximately three years in order to reassess the potential for further investments in capability. Ideally these investments would be guided through the delineation of grand challenge problems as prioritized by the biological research community.

We close this executive summary with some examples of activities which meet the criteria for grand challenges as discussed above. Past examples of such activities are the Human Genome Initiative and the design of an autonomous vehicle. It should be emphasized that our considerations below are by no means exhaustive. They are simply meant to provide example applications of a methodology that could lead to identification of such grand challenge problems and thus to a rationale for significant investment in high-performance capability or capacity. The possible grand challenges considered in our study were as follows:

1. The use of molecular biophysics to describe the complete dynamics of an important cellular structure, such as the ribosome;
2. Reconstructing the genome sequence of the common ancestor of placental mammals;
3. Detailed neural simulation of the retina;
4. The simulation of a complex cellular activity such as chemotaxis from a systems biology perspective.

We describe briefly some of the example challenges as well as their connection to opportunities for the application of advanced computation. Further details can be found in the full report.

A grand challenge that has as its goal the use of molecular biophysics to describe, for example, the dynamics of the ribosome would be to utilize our current understanding in this area to simulate, on biologically relevant time scales, the dynamics of the ribosome as it executes its cellular function of translation. The community of researchers in the area relevant to this grand challenge can be characterized as highly computationally-savvy and fully capable of effectively exploiting state-of-the-art capability. However, there remain significant challenges regarding the ability of current algorithms deployed on present-day massively parallel systems to yield results for time scales and length scales of true biological relevance. For this reason, significant investment in capability toward this type of grand challenge would, in our view, lead to only incremental gains given our current state of knowledge relevant to this problem. Instead, continuing investment is required in new algorithms in computational chemistry, novel computational architectures, and, perhaps most importantly, theoretical advances that overcome the challenges posed by the enormous range of length and time scales inherent in such a problem.

The second grand challenge considered by JASON is directed at large scale whole genome analysis of multiple species. The specific computational challenge is to reconstruct an approximation to the complete genome of the common ancestor of placental mammals, and determine the key changes that have occurred in the genomes of the present day species since their divergence from that common ancestor. This will require substantial computation for assembly and comparison of complete or nearly complete mammalian genomic sequences (approximately 3 billion bases each), development of more accurate quantitative models of the molecular evolution of whole genomes, and use of these models to optimally trace the evolutionary history of each nucleotide subsequence in the present day mammalian genomes back to a likely original sequence in the genome of the common placental ancestor. The computational requirements involve research in combinatorial algorithms, deployment of advanced high-performance shared memory computation as well as capacity computing in order to fill out the missing mammalian genomic data. A focused initiative in this area (or areas similar to this) in principle fulfills the JASON requirements for a grand challenge.

In the area of neurobiology, JASON considered the simulation of the retina as a potential grand challenge. Here a great deal of the fundamental functionality of the relevant cellular structures (rods, cones, bipolar and ganglion cells) is well established. There are roughly 130 million receptors in the retina but only 1 million optic nerve fibers, implying that the retina performs precomputation before processing by the brain via the optic nerve. Models for the various components have been developed and it is conceivable that the entire combined network structure could be simulated using today's capability platforms with acceptable processing times. Taken together, these attributes satisfy the requirements for a grand challenge, although it should be noted that current capability is probably sufficient for this task.

The final potential grand challenge considered in our study is the use of contemporary systems biology to simulate complex biological systems with mechanisms that are well-characterized experimentally. Systems biology attempts to elucidate specific signal transduction pathways and genetic circuits and then uses this information to map out the entire "circuit/wiring diagram" of a cell, with the ultimate goal of providing quantitative, predictive computational models connecting properties of molecular components to cellular behaviors. An important example would be the simulation of bacterial chemotaxis, where an enormous amount is currently understood about the cellular "parts list" and signaling network that is used to execute cellular locomotion. A simulation of chemotaxis that couples external stimuli to the signaling network would indeed be a candidate for advanced computational capability. At present, however, the utility of biological "circuits" as a descriptor of the system remains a topic for further research. Indeed, some recent experimental results indicate that a definite circuit topology is not necessarily predictive of system function. Further investigation is required to understand cellular signaling mechanisms before a large scale simulation of the locomotive behavior can be attempted. For this reason the chief impediment comes not from lack of adequate computing power, but from the need to understand better the signaling mechanisms of the cell.

2 INTRODUCTION

In this report we summarize the considerations and conclusions of the 2004 JASON summer study on high performance biocomputation. The charge to JASON (from DOE) was to

> "...explore the opportunities and challenges presented by applying advanced computational power and methodology to problems in the biological sciences... (JASON) will investigate the current suite of computationally intensive biological work, such as molecular modeling, protein folding, and database searches, as well as potential future endeavors (comprehensive multi-scale models, studies of systems of high complexity...). This study will also consider how advanced computing capability and capacity could best be brought to bear on bioscience problems, and will explore how different computing approaches (Grid techniques, supercomputers, commodity cluster computing, custom architectures...) map onto interesting biological problems."

The context for this study on high performance computation as applied to the biological sciences originates from a number of important developments:

- Achievements such as the Human Genome Project, which has had a profound impact both on biology and the allied areas of biocomputation and bioinformatics, making it possible to analyze sequence data from the entire human genome as well as the genomes of many other species. Important algorithms have been developed as a result of this effort, and computation has been essential in both the assimilation and analysis of these data.

- The DOE GTL initiative, which uses new genomic data from a variety of organisms combined with high-throughput technologies to study proteomics, regulatory gene networks and cellular signaling pathways, as well as more complex processes involving microbial communities. This initiative is also currently generating a wealth of data. This data is of intrinsic interest to biologists, but, in addition, the need to both organize and analyze these data is a current challenge in the area of bioinformatics.
- Terascale computation (meaning computation at the rate of $\approx 10^9$ operations per second and with storage at the level of $\approx 10^{12}$ bytes) has become increasingly available and is now commonly used to enable simulations of impressive scale in all areas of computational biology. Such levels of computation are not only available at centralized supercomputing facilities around the world, but are also becoming available at the research group level through the deployment of clusters assembled from commodity technology.

2.1 The Landscape of Computational Biology

The landscape of computational biology includes almost every level in the hierarchy of biological function, and thus the field of computational biology is almost as vast as biology itself. This is figuratively illustrated in Figure 2-1. Computation impacts the study of all the important components of this hierarchy:

1. It is central to the analysis of genomic sequence data where computational algorithms are used to assemble sequence from DNA fragments. An important example was the development of "whole genome shotgun sequencing" [20] which made it possible for Venter and his colleagues to rapidly obtain a rough draft of the human genome.

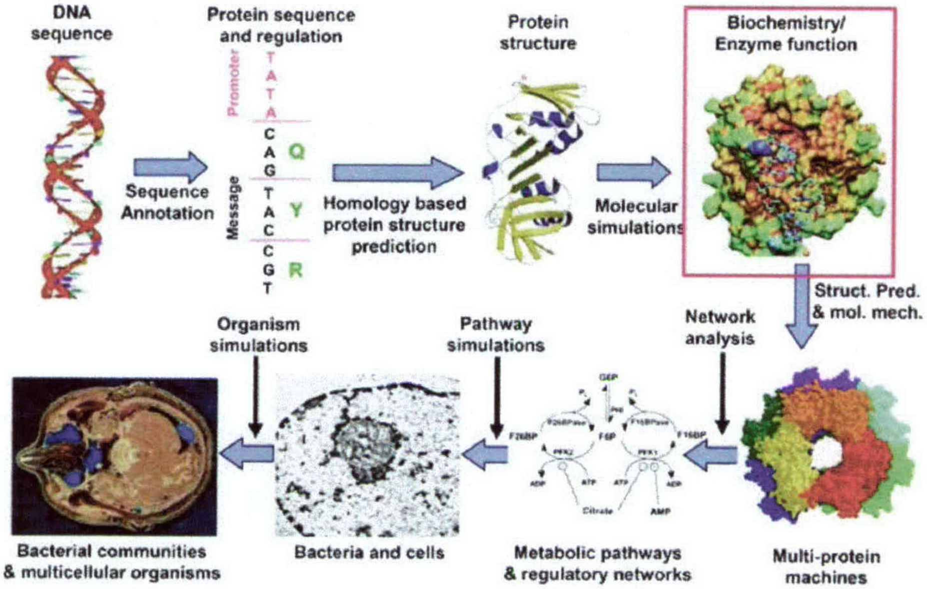
DNA sequence
Sequence Annotation
Protein sequence and regulation
Promoter
T
A
T
A
C
A
G
Q
Message
T
A
C
Y
C
G
T
R
Homology based protein structure prediction
Protein structure
Molecular simulations
Biochemistry/ Enzyme function
Struct. Pred. & mol. mech.
Network analysis
Pathway simulations
Organism simulations
Bacterial communities & multicellular organisms
Bacteria and cells
Metabolic pathways & regulatory networks
Multi-protein machines

Figure 2-1: A pictorial representation of the landscape of computational biology which includes almost every level in the hierarchy of biological function. Image from briefing of Dr. M. Colvin.

2. Via the processes of transcription and translation, DNA encodes for the set of RNAs and proteins required for cellular function. Here computation plays a role through the ongoing endeavor of annotation of genes which direct and regulate the set of functional macromolecules.

3. The function of a protein is tied not only to its amino acid sequence, but also to its folded structure. Here computation is essential in attempting to understand the relationship between sequence and fold. A variety of methods are applied ranging from so-called ab initio approaches using molecular dynamics and/or computational quantum chemistry to homology-based approaches which utilize comparisons with proteins with known folds. These problems continue to challenge the biocomputation research community.

4. Once the structure of a given protein is understood, it becomes important to understand its binding specificity and its role in cellular funct ion.

5. At a larger scale are cellular "machines" formed from sets of proteins which enable complex cellular activities. Simulation of these machines via computation can provide insight into cellular behavior and its regulation.

6. The regulation of various cellular machines is controlled via complex molecular networks. One of the central goals of the new area of "systems biology" is to quantify and ultimately simulate these networks.

7. The next levels comprise the study of cellular organisms such as bacteria and ultimately complex systems such as bacterial communities and multicellular organisms.

To cope with this vast landscape, the JASON study described in this report was focused on a selected set of topics where the role of computation is viewed as increasingly important. This report cannot be viewed therefore as exhaustive or encyclopedic. We note that an NRC report with much greater coverage of the field will be available in the near future [49]. During the period of June 28 through July 19, 2004 JASON heard briefings in the areas of

- Molecular biophysics
- Genomics
- Neural simulation
- Systems biology

These subfields are themselves quite large and so, again, our study represents a specific subset of topics. The complete list of briefers, their affiliations, and their topics can be found in the Appendix.

2.2 Character of Computational Resources

In assessing the type of investment to be made in computation in support of selected biological problems, it is important to match the problem under consideration to the appropriate architecture. In this section we very briefly outline the possible approaches. Broadly speaking we can distinguish two major approaches to deploying computational resources: capability computing and capacity computing.

Capability computing is distinguished by the need to maintain high arithmetic throughput as well as high memory bandwidth. Typically, this is accomplished via a large number of high performance compute nodes linked via a fast network. Capacity computing typically utilizes smaller configurations possibly linked via higher latency networks. For some tasks (e.g. embarrassingly parallel computations, where little or no communication is required), capacity computing is an effective approach. A recent extension of this idea is Grid computing, in which computational resources are treated much like a utility and are aggregated dynamically as needed (sometimes coupled to some data source or archive) to effect the desired analysis. The requirements as regards capability or capacity computing for biocomputation vary widely and depend to a large measure on the type of algorithms that are employed in the solution of a given problem and, in particular, on the arithmetic rate, memory latency and bandwidth required to implement these algorithms efficiently.

It is useful at this point to review the basic approaches in support of these requirements. We quote here the taxonomy of such machines as presented in the recent JASON report on the NNSA ASCI program [36]:

Custom: Custom systems are built from the ground-up for scientific computing. They use custom processors built specifically for scientific

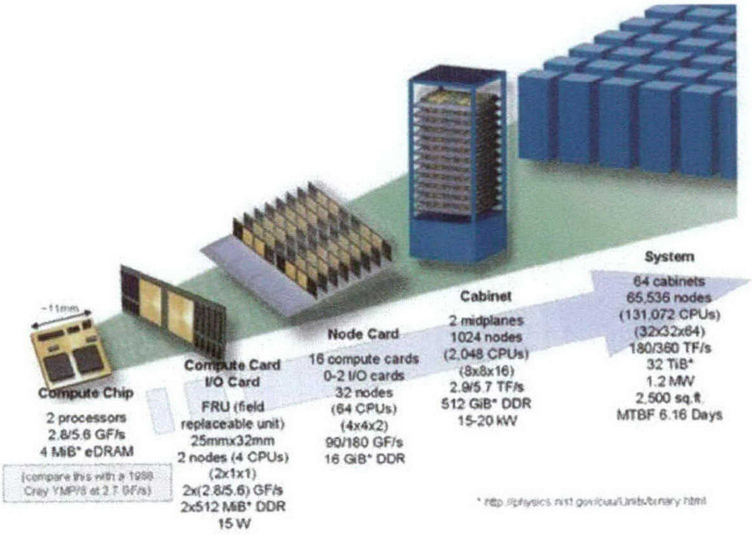

Figure 2-2: Hardware design schematic for IBM's Blue Gene/L.

computing and have memory and I/O systems specialized for scientific applications. These systems are characterized by high local memory bandwidth (typically 0.5 words/floating point operation (W/Flop), good performance on random (gather/scatter) memory references, the ability to tolerate memory latency by supporting a large number of outstanding memory references, and an interconnection network supporting inter-node memory references. Such systems typically sustain a large fraction (50%) of peak performance on many demanding applications. Because these systems are built in low volumes, custom systems are expensive in terms of dollars/peak Flops. However, they are typically more cost effective than cluster-based machines in terms of dollars/random memory bandwidth, and for some bandwidth-dominated applications in terms of dollars/sustained Flops. An example of custom architecture is IBM's recently introduced BlueGene computer. The architecture is illustrated in Figure 2-2. Such systems are typically viewed as capability systems.

Commodity-Cluster: Systems are built by combining inexpensive off-the-shelf workstations (e.g., based on Pentium 4 Xeon processors) using a third-party switch (e.g., Myrinet or Quadrics) interfaced as an I/O device. Because they are assembled from mass-produced components, such systems offer the lowest-cost in terms of dollars/peak Flops. However, because the inexpensive workstation processors used in these clusters have lower-performance memory systems, single-node performance on scientific applications suffers. Such machines often sustain only 0.5% to 10% of peak FLOPS on scientific applications, even on just a single node. The limited performance of the interconnect can further reduce peak performance on communication-intensive applications. These systems are widely used in deploying capacity computing.

SMP-Cluster: Systems are built by combining symmetric multi-processor (SMP) server machines with an interconnection network accessed as an I/O device. These systems are like the commodity-cluster systems but use more costly commercial server building blocks. A typical SMP node connects 4–16 server microprocessors (e.g., IBM Power 4 or Intel Itanium2) in a locally shared-memory configuration. Such a node has a memory system that is somewhat more capable than that of a commodity-cluster machine, but, because it is tuned for commercial workloads, it is not as well matched to scientific applications as custom machines. SMP clusters also tend to sustain only 0.5% to 10% peak FLOPS on scientific applications. Because SMP servers are significantly more expensive per processor than commodity workstations, SMP-cluster machines are more costly (about 5×) than commodity-cluster machines in terms of dollars/peak FLOPS. The SMP architecture is particularly well suited for algorithms with irregular memory access patterns (e.g., combinatorially based optimization methods). Small SMP systems are commonly deployed as capacity machines, while larger clusters are viewed as capability systems. It should be noted too that the programming model supported via SMP clusters, that is, a single address space, is considered the easiest to use in terms of the transformation of serial code to parallel code.

Hybrid: Hybrid systems are built using off-the-shelf high-end CPUs in combination with a chip set and system design specifically built for scientific computing. They are *hybrids* in the sense that they combine a commodity processor with a custom system. Examples include Red Storm that combines an AMD "SledgeHammer" processor with a Cray-designed system, and the Cray T3E that combined a DEC Alpha processor with a custom system design. A hybrid machine offers much of the performance of a custom machine at a cost comparable to an SMP-cluster machine. Because of the custom system design, a hybrid machine is slightly more expensive than an SMP-cluster machine in terms of dollars/peak FLOPS. However, because it leverages an off-the-shelf processor, a hybrid system is usually the most cost effective in terms of dollars/random memory band width and for many applications in terms of dollars/sustained FLOPS. Due to the use of custom networking technology and other custom features such systems are typically viewed as being capability systems.

2.3 Grand challenges

From the discussion in Section 2.1 it is not difficult to make a case for the importance of computation. However, our charge focused on the identification of specific opportunities where a significant investment of resources in computational capability or capacity could lead to significant progress. When faced with the evaluation of a scientific program and its future in this context, JASON sometimes turns to the notion of a "Grand Challenge". These challenges are meant to focus a field on a very difficult but imaginably achievable medium-term (ten-year) goal. Via these focus areas, the community can achieve consensus on how to surmount currently limiting technological issues and can bring to bear sufficient large-scale resources to overcome the hurdles. Examples of what may be viewed as successful grand challenges are the Human Genome Project, the landing of a man on the moon and, although, not yet accomplished, the successful navigation of an autonomous vehicle in the Mojave desert. Examples of what, in our view, are failed grand challenges include the "War on Cancer" (circa 1970) and the "Decade of the Brain" in which an NIH report in 1990 argued that neurobiological research was poised for a breakthrough, leading to the prevention, cure or alleviation of neurological disorders affecting vast numbers of people.

With the above examples in mind, JASON put forth a set of criteria to assess the appropriateness of a grand challenge for which a significant investment in high-performance computation (HPC) is called for. In the following sections of this report we then apply these criteria to various proposed grand challenges to assess the potential impact of HPC as applied to that area. It should be emphasized that our considerations below are by no means exhaustive. Instead, they are simply meant to provide example applications of a methodology that could lead to identification of such grand challenge problems and thus to a rationale for significant investment in high-performance capability or capacity.

The JASON criteria for grand challenges are

- A one-decade time scale: Everything changes much too quickly for a multi-decadal challenge to be meaningful.
- Grand challenges cannot be open-ended: It is not a grand challenge to "understand the brain", because it is never quite clear when one is done. It is a grand challenge to create an autonomous vehicle that can navigate a course that is unknown in advance without crashing.
- One must be able to see one's way, albeit dimly, to a solution. When the Human Genome Project was initiated, it was fairly clear that it was, in principle, doable. The major issue involved improving sequencing throughput and using computation (with appropriate fast algorithms) to facilitate the assembly of sequence reads. While underscoring the

tremendous importance of these advances, they are not akin to true scientific breakthroughs. Thus, one could not have created a grand challenge to understand the genetic basis of specific diseases in 1950 before the discovery of the genetic code. This is independent of how much data one might gather on inheritance patterns, etc. With some important exceptions, data cannot, in general, be back-propagated to a predictive "microscopic" model. One must therefore view with some caution the notion that we will enter a data-driven era when scientific hypotheses and model building will become passé.

- Grand challenges must be expected to leave an important legacy. While we sometimes trivialize the space program with jokes about drinking Tang, the space program did lead to many important technological advances. This goes without saying for the human genome project. This criteria attempts to discriminate against one-time stunts.

The remaining sections of this report provide brief overviews of the role of computation in the four areas listed in section 2.3. At the end of each section we consider possible grand challenges. Where a grand challenge seems feasible we describe briefly the level of investment of resources that would be required in order to facilitate further progress. Where we feel the criteria of a grand challenge are not satisfied we attempt to identify the type of investment (e.g. better data, faster algorithms, etc.) that would enable further progress.

3 MOLECULAR BIOPHYSICS

Molecular biophysics is the study of the fundamental molecular constituents of biological systems (proteins, nuclei acids and specific lipids) and their interaction with either small molecules or each other or both. These constituents and their interactions are at the base of biological functionality, including metabolism, gene expression, cell-cell communication and environmental sensing, and mechanical/chemical response. Reasons for studying molecular biophysics include:

1. The design of new drugs, enabled by a quantitatively predictive capability in the area of ligand-binding and concomitant conformational dynamics.

2. The design and proper interpretation of more powerful experimental techniques. We briefly discuss in this section the role of computation in image analysis for biomolecular structure, but this is only one aspect of this issue[2].

3. A better understanding of the components involved in biological networks. Current thinking in the area of systems biology posits that one can think of processes such as genetic regulatory networks as akin to electrical circuits[3]. The goal here is to find the large scale behavior of these networks. But recent experiments have provided evidence that this claim, that we know enough of the constituents and their interactions to proceed to network modeling, may be somewhat premature.

[2] A notable development discussed during our briefings was a recent case where quantum chemistry calculations helped in the design of a green fluorescent protein (GFP) fusion, in which attaching GFP to a functional protein and carefully arranging the interaction led to the capability of detecting changes in the conformational state of the protein – these probes will offer a new window on intra-cellular signaling, as information is often transmitted by specific changes (such a phosphorylation) in proteins, not merely by their presence or absence.

[3] This metaphor is responsible for attempts to create programs such as BioSpice, modeled after the SPICE program for electrical circuits

Issues such as the role of stochasticity, the use of spatial localization to prevent cross-talk, the context-dependent logic of transcripts factors etc. must be addressed via a collaboration of molecular biophysics and systems biology. Further discussion of these issues can be found in Section 6.

4. Development of insight into the unique challenges and opportunities faced by machines at the nano-scale. As we endeavor to understand how biomolecular complexes do the things they do, undeterred by the noisy world in which they live, we will advance our ability to design artificial nano-machines for a variety of purposes.

In the following, we will briefly survey three particular research areas in which computation is a key component. These are imaging, protein folding, and biomolecular machines. We will see specific instantiations of the aforementioned general picture. We then consider a possible grand challenge related to this area - the simulation of the ribosome.

3.1 Imaging of Biomolecular Structure

One of the areas where computational approaches are having a large effect is in the development of more powerful imaging techniques. We heard from W. Chiu (Baylor College of Medicine) about the specific example of the imaging of viral particles by electron microscopy. Essentially, a large number of different images (i.e. from different viewing perspectives) can be merged together to create a high resolution product. To get some idea of the needed computation, we focus on a 6.8 Å structure of the rice dwarf virus. This required assembling 10,000 images and a total computation time of $\sim$ 1500 hours on a 30 CPU Athlon (1.5 GHz) cluster (a very conventional cluster from the viewpoint of HPC). This computation is data-intensive but has modest memory requirements (2 GByte RAM per node is sufficient).

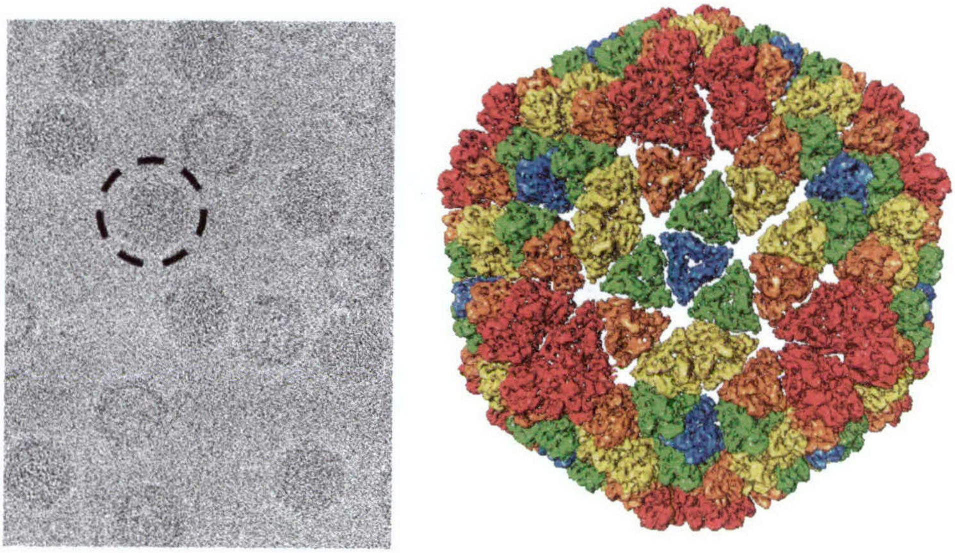

Figure 3-3: An image of the outer capsid of the Rice Dwarf Virus obtained using cryo-electron microscopy. The image originates from a briefing of Dr. Wah Chiu.

A typical result is shown in Figure 3-3. Remarkably, the accuracy is high enough that one can begin to detect the secondary structure of the viral coat proteins. This is facilitated by a software package developed by the Chiu group called Helix-Finder, with results shown in Figure 3-4. The results have been validated through independent crystallography of the capsid proteins. One of the interesting questions one can ask relates to how the computing resource needs scale as one moves to higher resolution. Dr. Chiu provided us with estimates that 4Å resolution would require 100,000 images and about 10,000 hours on their existing small cluster. If one imagines a cluster which is ten times more powerful, the image reconstruction will require a year's worth of computation as this is an embarrassingly parallel task. This is enough to put us (marginally) in the HPC ball park, but there is no threshold here – the problem can be done almost equally well on a commodity cluster, or potentially via the Grid, and this will lead to only a modest degradation in the resolution achievable by a truly high-end machine. Because the type of image reconstruction as described by Dr. Chiu is an embarrassingly parallel

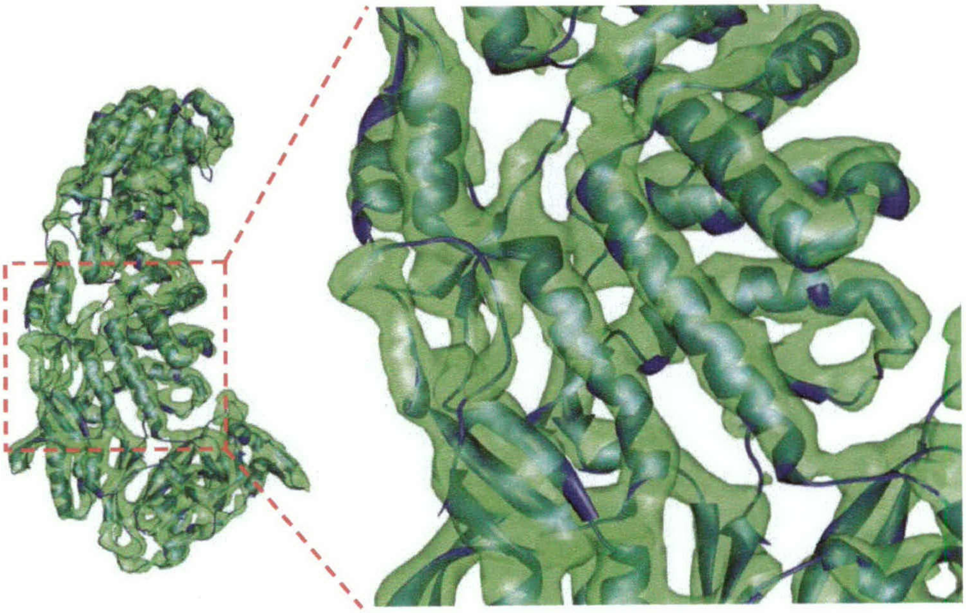

Figure 3-4: Identification of helical structure in the outer coat proteins of the rice dwarf virus. Image from briefing of Dr. Wah Chiu (Baylor College of Medicine.

computation, one can make a cogent argument for deployment of capacity computing and, indeed, the development of on-demand network computing, a signature feature of Grid computing, would be a highly appropriate approach in this area.

Imaging in biological systems is a field which certainly transcends the molecular scale; its greatest challenges are at larger scales where the concerted action of many components combine to create function. These topics are not part of molecular biophysics and so are not discussed here. For some more information one can consult a recent JASON report [39] on this topic.

3.2 Large-scale molecular-based simulations in biology

We next assess several aspects of molecular-based simulation that are relevant to high performance computation. There has been major progress in molecular scale simulations in biology (i.e., including biophysics, biochemistry, and molecular biology) since the first molecular dynamics (MD) calculations from the early 1970's. The field has evolved significantly with advances in theory, algorithms, and computational capability/capacity.

Simulations include a broad range of energetic calculations that include MD, Monte Carlo methods (both classical and quantum), atomic/electronic structure dynamics optimization, and other statistical approaches. In MD, the trajectories of the component particles are calculated by integrating the classical equations of motion. The simplest renditions are based on classical force fields that use parameters (e.g., force constants) derived from fitting to experimental data or to theoretical (quantum mechanical) calculations. These can be supplemented by explicit quantum mechanical calculations of critical components of the system [45, 14, 26]. These calculations are particularly important for modeling chemical reactions (i.e., making and breaking bonds). At the other end of the scale are continuum approaches that ignore the existence of molecules. In fact, it has been fashionable to use hybrid approaches involving quantum mechanical, classical molecular, and continuum methods to model the largest systems. In addition to the intrinsic accuracy problems with each of the component parts (discussed below), there are important issues on how to appropriately describe and treat the interfaces between the quantum, classical, and continuum regimes [40].

It is a truism from physics that a full quantum mechanical treatment of a biological system would yield all necessary information required to explain its function if such a treatment were tractable [10]. The reality of course, is that existing methods for the quantum mechanical treatment of even a small piece of the problem (e.g. the active site of an enzyme) are still approximate and the accuracy of those methods needs to be carefully examined in the context of the problem that one is trying to solve. Some feel for the size of the problem can be obtained from Figure 3-5 where typical simulation approaches for molecular biophysics are put in context. As can be seen from the Figure, the applicability of a given method is linked to the number of

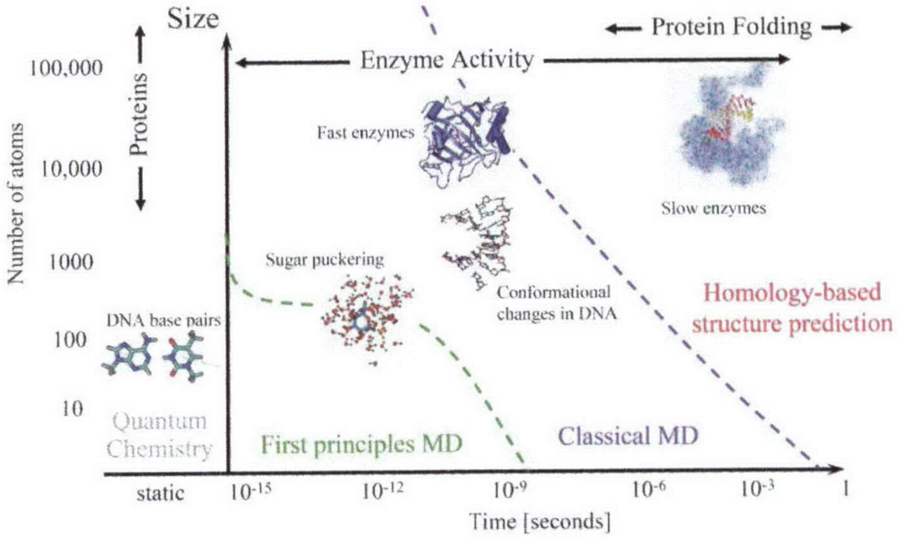

Figure 3-5: A plot of typical simulation methodologies for molecular biophysics plotted vs number of atoms and the relevant time scale. Figure from presentation of Dr. M. Colvin.

atoms in the system under consideration as well as the required time scale for the simulation. The larger the number of atoms or the longer is the required simulation time, the further one moves away from "ab initio" methods.

Quantum approaches break down into either so-called quantum chemical (orbital) or density functional methods. The quantum chemical methods have intrinsic limitations in terms of the number of electrons that can be simulated and the trade-off in basis set size versus system size impacts accuracy. The most accurate methods (including configuration interaction or coupled cluster approaches) typically scale as N^5 to N^7, where N is the number of electrons in the system. As a result of these limitations, there has been increasing interest in the use of density functional methods [23, 40], which have been used extensively in the condensed-matter physics community because of their reasonable accuracy in reproducing the ground-state properties of many semiconductors and metals. Despite the name "first-principles", there is an arbitrariness in the choice of density functionals (e.g. to model exchange-correlations) and there has been extensive effort to extend the local density approximation (e.g., with gradient corrections) or using other alternatives such as Gaussian Wave bases (e.g. [16]). While these extensions may more accurately represent the physics of the problem, the extensions can result in poorer agreement between theory and experiment. Following the Born-Oppenheimer approximation, the dynamics is treated separately from the forces (i.e. using the Hellman-Feynman theorem) and usually in the quasi-harmonic approximation.

The advent of first-principles MD has been an important breakthrough [7] and is being applied to a range of chemical and even biological problems [40]. Here the electronic structure is calculated on the fly as the nuclei move (classically), with the coefficients of the single-particle wave functions treated as fictitious zero-mass particles in the Lagrangian. The much larger size of the simulation relative to the classical case results in limitations to the basis set convergence, k-point sampling, choice of pseudopotentials, and system size. Moreover, these techniques are still based on density functional approximations, so the problems discussed above apply here as well. Because of this, the accuracy needs to be carefully examined. There are a number of problems to be surmounted before these methods can be fully implemented for biological systems (cf. for example [26, 3]). A full ab initio calculation of a small protein has been reported (1PNH, a scorpion toxin with 31 residues and 500 atoms; [3]). Hybrid classical and first-principles MD calculations have also been applied to heme [35].

One can step back from the problem of treating biomolecules, by considering the problem of accurately describing and calculating the most abundant molecules in biological systems: water. After years of effort, the proper treatment of water in condensed phase is still challenging. The most accurate representations of the physical properties of the molecule (i.e., with the proper polarizability) in condensed phase and in contact with solutes is often too time consuming to compute, so simple models are used. Indeed, the full first-principles approaches still fail to reproduce the important physical and chemical properties of bulk H_2O [17]. Studies of aqueous interface phenomena with these techniques are really only beginning [8].

In principle, the most accurate methods would be those that take the full quantum mechanical problem, treating the electrons and atoms on the same quantum mechanical footing. Such methods are statistical (e.g., various formulations of quantum Monte Carlo) or use path integral formulations for the nuclei [15]. In quantum Monte Carlo, the problem scales as N^3. Because of this, the treatment of heavy atoms (beyond H or He) has generally been problematic. But there are also fundamental problems. In the case of quantum Monte Carlo there is the fermion sign problem. Linear scaling methods have been developed so that systems of up to 1000 electrons can be treated (e.g., Fullerene [48]). These methods have not been applied directly to biomolecular systems to our knowledge.

Several additional points need to be made. The first is that biological function at the molecular level spans a broad range of time scales, from femtosecond scale electronic motion to milliseconds if not seconds. Independently of the intrinsic accuracy of the calculations (from the standpoint of energetics), the time-scale problem is beyond conventional molecular-based simulations. On the other hand, stochastic methods can bridge the gap between some time scales (i.e., molecular vibrations, reaction trajectories and large scale macromolecular motion [50, 13, 38]). This is also important for the protein folding problem [44]. Finally, the above discussion concentrates on the use of simulations for advancing our understanding of biological function from the standpoint of theory, essentially independent of experiment. On the other hand, there is a growing need for large-scale molecular-based simulations as an integral part of the analysis of experimental data. Classical MD and Monte Carlo (including reverse Monte Carlo) simulations can be used in interpreting data from diffraction, NMR, and other kinds of spectroscopy experiments [3]. These examples include chemical dynamics experiments carried out with sub-picosecond synchrotron x-ray sources. The needs here for high-performance computing appear to be significant. The computational chemistry community, however, has been very successful in articulating these requirements and will be able to make a cogent case for future resources required to support this work. The above discussion also underscores once again the need for basic research that can then lead to future consideration of larger systems of biological interest.

In order to provide some context for the scale of applications that one envisions, we close this section with a brief discussion of the computational resources required for a very basic protein folding calculation using a simple and conventional classical MD approach. In order to try to capture the interatomic interactions, use is typically made of various potentials with adjustable parameters that are used to fit data acquired from more accurate calculations on smaller systems. A typical set of such potentials (quoted from [2]) is expressed below:

$$
\begin{aligned}
U_{Total} &= U_{Stretch} + U_{Bend} + U_{Torsion} + U_{LJ} + U_{Coulomb} \qquad \text{where} \\
U_{Stretch} &= \sum_{bonds(ij)} K_{ij} \left(r_{ij} - r_{ij}^{equil} \right)^2 \\
U_{Torsion} &= \sum_{torsions(ijkl)} \sum_{n=1,2,\ldots} V_{ijkn} \left[1 + \cos(n\phi_{ijkl} - \gamma_{ijkl}) \right] \\
U_{LJ} &= \sum_{nonbonded(ij), i<j} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \\
U_{Coulomb} &= \sum_{nonbonded(ij)} \frac{q_{ij}}{\epsilon r_{ij}}
\end{aligned}
\tag{3-1}
$$

The total interaction is comprised of bonded and nonbonded interactions. The bonded interactions account for bending, stretch and torsion in the protein structure. Nonbonded interactions account for the electrostatic as well as Lennard-Jones interactions. Equation 3-1 represents the forces typically taken into account in MD simulations of protein and water systems. The accuracy of this expression is directly related to how the choice of the parameters (for example interaction strengths such as K_{ij}) is made. It is here that more accurate quantum chemical approaches might be used to create a valid "force field". The MD approach simply computes all the forces on all atoms of the protein (and solvent) and then adjusts the positions of the atoms in response to these forces. However, several aspects of this calculation are extremely challenging. They are summarized in the table below:

Physical time for simulation	10^{-4} seconds
Typical time step size	10^{-15} seconds
Typical number of MD steps	10^{11} steps
Atoms in a typical protein and water simulation	32000 atoms
Approximate number of interactions in force calculation	10^{9} interactions
Machine instructions per force calculation	1000 instructions
Total number of machine instructions	10^{23} instructions

The estimates come from [2]. As shown in the table, a typical desired simulation time might be on the order of 10-100 microseconds although it is known that folding timescales can be on the order of milliseconds or longer. The second entry illustrates one of the most severe challenges: in the absence of any implicit time integration approach the integration must capture the vibrational time scales of the system which are in the femtosecond range. The number of interactions required in the force calculation is derived from the most simple estimate wherein all $O(N^2)$ interactions are computed for a system of size N. This can be in principle be reduced through the use of methods based on multipole expansions; this entails significant programming complexity when one contemplates implementing such algorithms on parallel architectures and improvement over the simple approach will not be seen until N is sufficiently large. As a result the estimate provided above is probably not far off. In total, such folding calculations require 10^{23} operations to compute one trajectory. For a computer capable of a Petaflop such a calculation will still require $O(10^8)$ seconds or roughly three years.

A computer capable of arithmetic rates of a Petaflop is today only feasible through the use of massive parallelism. It is envisioned that computers capable of peak speeds of roughly a Petaflop will be available in the next few years. An example is the recently announced BlueGene/L machine from IBM which represents today the ultimate capability platform. The largest

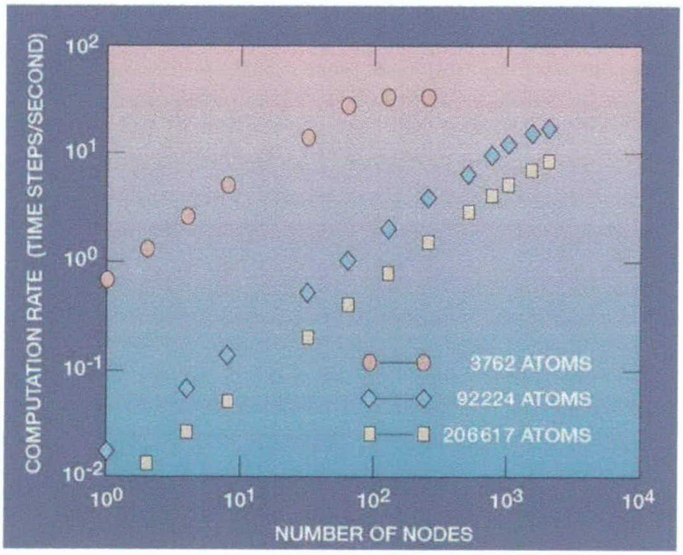

Figure 3-6: Scaling for the molecular dynamics code NAMD on the ASCI Red massively parallel computer.

configuration of this machine is 64000 processors each capable of roughly 3 Gflops. Thus, present-day configurations of BlueGene are capable of peak speeds of roughly .2 Petaflop and it is anticipated that through improvements in processor technology it will be possible to achieve peak speeds of a Petaflop in the very near future.

However, as discussed in section 2.2, it is difficult to achieve the ideal peak speed on a single processor. This is typically because of the inability to keep the processor's arithmetic units busy every clock cycle. Even without massive parallelism processors will perform at perhaps 0.5 to 10% of their peak capabilities. Further latency results when one factors in the need to communicate across the computer network. Communication is typically quite a bit slower than computation even in capability systems and so for some algorithms there can be issues of scalability as the number of processors are increased. Computations such as those required for protein folding exhibit significant nonlocality in terms of memory access and so the development of scalable algorithms is crucial. A example of this (based on rather old data) is shown in Figure 3-6. The figure shows the number of time steps that can be completed in one second of wall clock time using a modern protein simulation code NAMD. The data come from the ASCI Red platform which is now almost 10 years old. Nevertheless the trends are reflective of what can be expected to happen on more modern platforms. It can be seen that as the number of atoms is held fixed and the number of processors increased. the computational rate eventually saturates implying the existence of a point of diminishing returns. The performance can be improved by increasing the number of atoms per processor or by reducing network latency.

To conclude, it is seen that the computational requirements for highly accurate molecular biophysics computations are significant. The challenge of long time integration is particularly severe. We discuss in more detail the particular problem of protein folding in the next section.

3.3 Protein Folding

One of the most computation-limited problems currently being vigorously pursued is that of protein folding. Actually, there are two separate folding problems that should not always be lumped together. The first is the determination of protein structure from the underlying amino acid sequence; there is a corresponding nucleic acid problem of determining the structure of single-stranded RNA from nucleotide sequence. This problem has its final goal an atomic level picture of the folded-state conformation but does not necessarily care about the folding kinetics. The second problem is the time evolution of protein folding, determining the exact set of trajectories that enable the system to fold from an initially unfolded ensemble. Here one cares about the folding kinetics and the nature of the transition states. This information can be crucial, as in for example the problem of protein aggregation disease due to the clumping together of proteins that have gotten trapped in misfolded non-native states.

The "holy grail" in this field is being able to directly simulate the folding process of a typically-sized single domain protein (say 100 residues) starting from a random initial state. This would presumably be done by classical MD with a well-established force field and in the presence of explicit water molecules (i.e. solvation). This simulation would of course directly address the structure problem and would demonstrate at least one kinetic trajectory; presumably multiple runs would be needed to determine the full range of possible kinetic pathways. A first step towards the direct realization of this capability was made by Duan and Kollman [11], who simulated the folding of the 36 residue Villin head piece (see Figure 3-7) for one microsecond. The Villin head piece subdomain that was simulated is one of the most rapidly folding proteins and this calculation represented a new level of capability in this area. To give some idea of the resources required for these studies, their computation ran for several months on a 256 node cluster at the Pittsburgh Supercomputer Center. Despite the impressive scale of this type of computations there is, in our opinion, no compelling argument that brute force calculations enabled by a state-of-the-art capability machine are really going to break open this problem. It is not as if there is a well-defined force field that will give us the correct answer every time if only we had enough cycles. Such an approach is valid in some other fields (e.g. computational hydrodynamics, lattice quantum chromodynamics, etc.) but appears wholly inappropriate here as pointed out earlier in Section 3.2. Instead, the refinement of force fields must go hand in hand with a broad spectrum of computational experiments as discussed in the previous section. Furthermore, there is no one unique protein of interest and it is quite likely that models that suffice for one protein of one class will need to be modified when faced with a different structural motif – this has been seen when standard programs such as CHARMM and AMBER, usually calibrated on proteins that have significant α-helix secondary structure, are used for mostly β-sheet structures. The fact that one model will not do for all proteins is a consequence of assuming that the problem can be addressed by classical MD with few-body potentials. This is only approximately true as previously discussed in

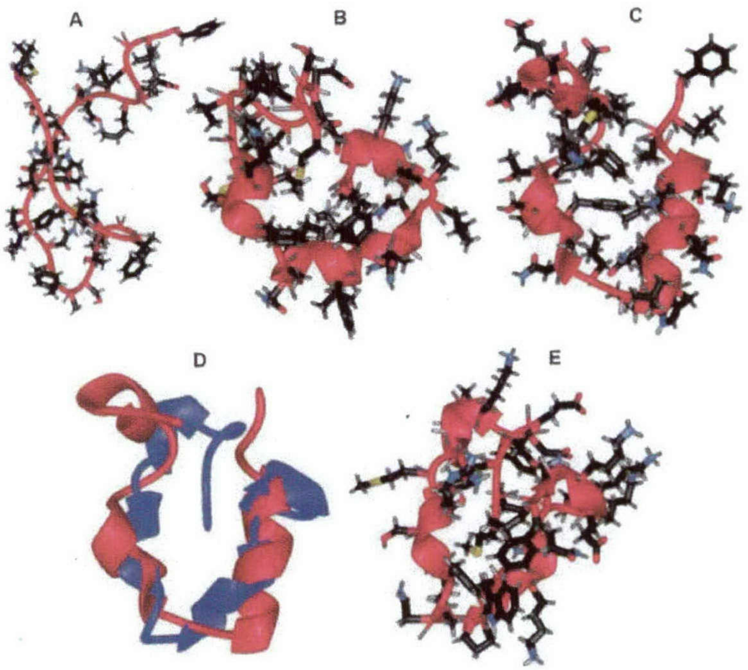

Figure 3-7: Representations of various stages of folding of the 36 residue Villin head piece. (A) represents the unfolded state; (B) a partially folded state at 980 nsec and (C) a native structure. (E) is a representative structure of the most stable cluster. (D) is an overlap of the native (red) and the most stable cluster (blue) structures indicating the level of correlation achieved between the simulation and a native fold. (Figure from [11]).

section 3.2; the real interactions are quantum mechanical and many-body in nature, and hence empirical adjustments must be made on a case-by-case basis. Of course, the idea of going beyond classical MD to a more "realistic" ab initio treatment (using density functional theory, for example) would appear to be totally out of the question using present computational techniques given the considerations discussed in section 3.2

Even in the absence of a direct path to the answer, the molecular biophysics community continues to make excellent progress by using a variety of approximations, simplifying assumptions and, of primary concern here, computational resources and paradigms. It is not useful to give a comprehensive review, but it is worth presenting some of the highlights of these alternate

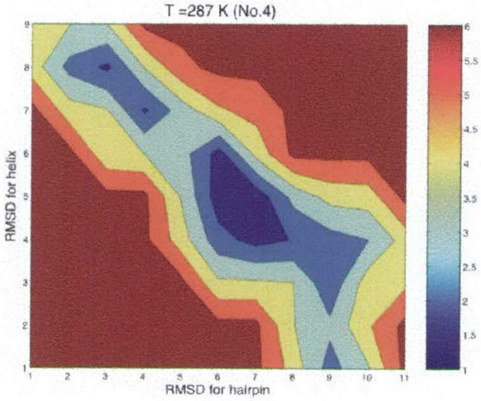

Figure 3-8: A free energy plot of a 16 residue segment of some specific protein; the axes refer to distance away from an alpha helix (y-axis) versus a beta hairpin (x-axis), and blue is low free energy (i.e. a probable configuration). Here, the number of atoms being simulated is 8569 and the calculation is done using 42 replicas.

approaches:

Simplified models: If one retreats from all-atom simulations, one can get longer runs of the folding time-course. One can eliminate the water molecules (going to "implicit solvent" models), eliminate the side chains (so-called C^{α} models) and even restrict the overall conformational space by putting the system on a lattice. These have been used to great effect to study folding kinetics. These simulations run quite effectively on existing clusters which have become the standard resource for the community.

Thermodynamics: If one is willing to give up on folding kinetics and merely study the thermodynamics of the system, advanced sampling techniques enable rapid exploration of the conformational space. For example, the replica exchange method uses a set of replicas that evolve at differing temperatures. Every so often, configurations are swapped between replicas, preventing the low temperature system of interest from getting trapped for long periods of time. Because of limited communication between the replicas, this algorithm is close to being embarrassingly parallel. As an example, we show in Figure 3-8 the free energy plot of a 16 residue segment of some specific protein; the axes refer to distance away from an α-helix (y-axis) versus a β-hairpin (x-axis), and blue is low free energy (i.e. a probable configuration). Here, the number of atoms being simulated is 8569 and the calculation is done using 42 replicas. These data are based on a 6 nanosecond MD simulation, which took 96 hours on the SDSC Teragrid machine with 168 CPU's.

Folded State: If one is interested only in the native state, one can dispense with kinetics altogether and focus on finding the minimal energy structures. This can be tackled by a whole set of possible optimization algorithms. Many of the practitioners of these techniques compete in the CASP competition to predict structures which have been measured but as yet not-released. As we heard from one of our briefers, Peter Wolynes, progress is being made on structure prediction by "folding in" theoretical ideas such as the relative simplicity of the energy landscape for natural proteins.

Grid-based methods Several groups are exploring the distributed computing paradigm for performing folding computations. One interesting idea is due to Pande [33] who noted that for simple two-state folding kinetics, the folding is a Poisson process (i.e. has an exponential waiting time distribution). This means that one can run thousands of totally independent folding simulations and that a small percentage ($\sim t/t_{\text{folding}}$) will fold after a small time t. They have demonstrated how this simplifying assumption can be used to harness unused computational capacity on the Internet to actually get folding paths. Other groups are also beginning to explore distributed computer applications (see, for example, the work of the Brooks group [6] at Scripps Research on structure determination for the CASP6 competition). These application are being facilitated by the increasing availability of Grid middleware (cf., for example the Berkeley Open Infrastructure for Network Computing project [1]).

We should mention in passing that most of the work to date has focused on soluble proteins. An additional layer of complexity occurs when proteins interact directly with membranes, such as for example when the parts of the protein repeatedly traverse the lipid bilayer. Additional attention is being paid to this topic, but progress remains sporadic, especially since the structural data upon which the rest of the folding field is directly reliant, is much harder to come by.

In summary, the protein folding problem will use up all the cycles it can and will do so with good effect. Progress is being made by using a whole suite of computational platforms together with theoretical ideas which motivate simplifying assumptions and thereby reduce the raw power needed. This mix appears to us to be the most promising direction; a single dedicated facility for protein folding (as was advertised initially for Blue Gene) will be useful but would not on its own break the field open. We elaborate on this issue further in the next section.

3.4 A Potential Grand Challenge - The Digital Ribosome

The understanding of biomolecular structure, while clearly important in its own right, is but a step towards the more essential area of biomolecular function, that is, how the dynamic three dimensional structure of biomolecules and biomolecular complexes enable critical steps in the life-cycle of organisms to be carried out. The simplest of these possibilities is the catalyzing of a specific reaction by a single component enzyme; other "simple" functions include the capture of a photon by a light-sensitive pigment embedded in a protein photoreceptor. More complex possibilities include the transduction of chemical energy into mechanical work, the transfer of molecules across membranes, and the transfer of information via signaling cascades (often with the use of scaffolds for spatial organization of the reactions). At the high end of complexity one has incredibly intricate multi-component machines which undergo large scale conformational motions as they undertake tasks. A classic example is the ribosome, consisting of roughly 50 proteins and associated RNA molecules. Its job, of course, is to translate the sequence of messenger RNA into the amino acid sequence of a manufactured protein.

Typically, studies of biomolecular function of the underlying structure are advanced via X-ray crystallography, cryo-electron microscopy or NMR. Often, one can obtain several static pictures of the complex, perhaps with bound versus unbound ligand for example. The challenge is then to understand the sequence of events comprising the actual functioning of the machine. The complexity arises from the need to keep track of a huge number of atoms (millions, for the ribosome) and from the need to do some sort of quantum mechanical treatment of any of the actual chemical reactions taking place.

Let us again focus on the quantum chemistry aspects of the problem (as discussed in Section 3.2). It is clear that one cannot hope to do justice to any of the quantum aspects of the problem for more than a tiny fraction of the biomolecular complex, and for more than a tiny fraction of the time involved an entire functional cycle. This part of the problem has to then be coupled to the rest of the dynamics in space and time which presumably are being treated by classical MD simulations. This task falls under the general heading of "multi-scale computation" where part of the problem needs to be done at a very much finer resolution than others. Our impression is that there remains much room for algorithmic improvement for this interfacing task. We heard about progress on quantum algorithms from various briefers. This community is rather mature and is making steady progress, but again, it did not appear from our briefings that deployment of HPC would at this point create a "sea-change" in our current understanding of biological function. Instead, we see a mix of platforms being applied in valuable ways to various problems with achievement of incremental progress.

The biggest problem in this area appears to be the "serial time" bottleneck. HPC can, in principle, allow us to consider bigger systems although there are issues of scalability, but cannot directly address the difficulty of integrating longer in time if one uses conventional "synchronous" integration methods. The mismatch in time scale between the fundamental step in a dynamical simulation (on the order of femtoseconds) and the time-scale of the desired functional motions (milliseconds or longer) is depressingly large and will remain the biggest problem for the foreseeable future. Of course, there are ways to make progress. One familiar trick is driving the system so hard that the dynamics speeds up; the extent to which these artificial motions are similar to the ones of interest needs to be carefully investigated on a case by case basis. Finding some analytic method which allows for integrating out rapid degrees of freedom is obviously something to aim at, but again any proposed method should be carefully evaluated.

Within the context of biological machines, we consider the notion of the "Digital Ribosome" as a possible grand challenge in computational biology. Exactly how uniquely important the ribosome is as compared to other critical biological machines is somewhat subjective, but it is fair to say that it does represent a first-order intellectual challenge for the biology community. Namely, one wants to understand how the structure allows for the function, what is the purpose of all the layers of complexity, which pieces are the most constrained (and hence have the hardest time changing from species to species over evolutionary time) and, of course how did the ribosome (with all the attendant implications for cell biology) come to be. This problem has come to the fore mostly because of the remarkable recent successes in imaging of ribosomal structure (see for example Figure 3-9). The existence of structural information as well as the long history of using ribosomal RNA to track evolution seems to allow us to converge to a set of coherent tasks that

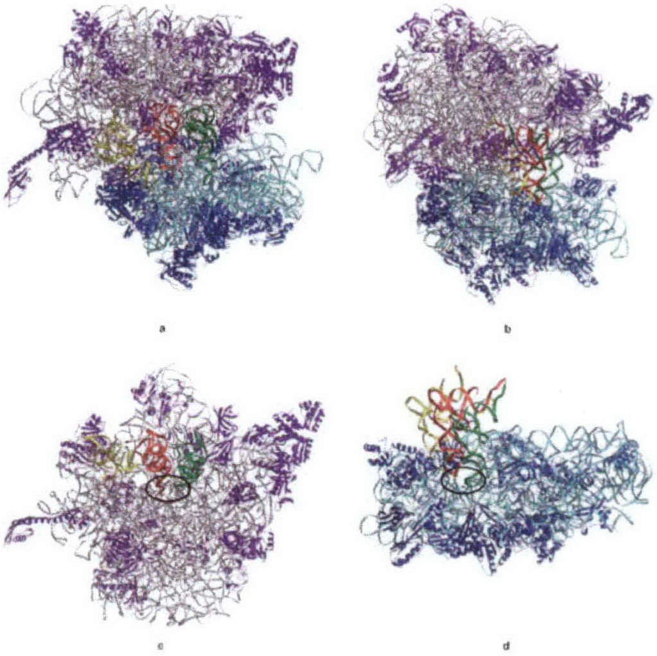

Figure 3-9: Views of the three dimensional structure of the ribosome including three bound tRNA's. (a) and (b) Two views of the ribosome bound to the three tRNAs. (c) The isolated 50S subunit bound to tRNAs - peptidyl transfer center is circled (d) Isolated 30S subunit bound to tRNAs- the decoding center is circled. The figure above is taken from [46]

would enable us to formulate this challenge. This would have a high payoff, one of our challenge criteria, and would actually energize the community. But, is it doable?

Our basic conclusion is that, at present, the serial bottleneck problem as well as our lack of fully understanding how to create classical force fields (as well as understanding when one needs to go to full ab initio methods) makes the digital ribosome project premature. We do not see a path to full simulation capability and, although there are promising approximate methods based on a sort of normal mode analysis, we do not yet understand how to do reliable dynamics without such a capability. This is only a weak conclusion, however, and we think that this issue should perhaps be put to the molecular biophysics community in a more direct fashion. Further, it is our opinion that the total decoupling of molecular biophysics calculations from evolutionary information is possibly holding back progress. After all, one can get some ideas of the most relevant residues by using comparative genomics and conversely one can make better sense of the variations observed in the ribosome in different species in"tree of life" if one has some handle on the functional robustness of the protein structure via direct calculations. Again, this underscores that progress can be made by coupling highly targeted and smaller scale computations with experimental information.

3.5 Conclusion

In the course of our study, we heard briefings from many different areas of computational biology. It was clear that the area of molecular biophysics is the most computationally sophisticated, the field in which computational methods have become of age. In areas ranging from the computer-aided analysis of advanced imaging methods to medium-scale solution of model equations to full-up simulations of the equations of motions for all the atoms using high performance computing assets, this field is moving forward and making impressive gains. So, there is every reason to continue work on the challenges facing this field. As we heard from our briefers and as we thought through the issues among ourselves, our primary question related to computation was one of investment strategy. Simply put, what mix of computational resources provides the best fit to today's research community and conversely, how would investment in high performance computing impact the progress to be made in the future?

Our basic conclusion is that an effective model for computational resource needs is an approach currently adopted by Klaus Schulten (Univ. Illinois) of attempting to provide a cluster per graduate student. In his lab, each student is given complete control of a commodity cluster (roughly 50 processors) for his/her research. Similarly, we heard from Dr. Chiu that clusters of this scale are the right tool for current imaging applications. The logic behind this is that

- there are many important problems to be worked on, not a single unique challenge (contrast this to QCD, for example).
- almost all problems require significant computation. There is a sort of "minimum complexity principle" at work, which means that even the simplest biologically meaningful systems are much more complex than most physicists care to admit. This tips the balance of simple soluble models/intermediate models requiring some simulation/detailed models requiring significant computation to the right of what is standard in most basic physics areas. A single workstation is clearly inadequate.
- We are far away from any very specific "threshold" of understanding. Our understanding of specific systems will continue to increase incrementally and no one set of "super-calculations" doable in the foreseeable future will have a first order effect on the field. Thus, there is limited utility in providing a very small number of researchers access to more computational cycles in the form of a HPC capability machine - this type of machine would be effectively utilized, but would probably not lead to breakthrough results.
- Conversely, there could be breakthroughs based either on algorithmic improvements or conceptual advances. One might argue, for example, that the idea of a "funneled landscape" (discussed above in 3.3) has led to useful simplified models and indeed to constraints on "realistic models" which have enhanced our ability to predict protein structure. New ideas for electrostatic calculations might fit into this category. These algorithms and/or ideas will only come from having many researchers trying many things, another argument for capacity over capability.

We comment here at this point on the deployment of software. We were struck by the fact that this community is quite advanced when it came to developing and maintaining useful software packages which can then be shared worldwide. These packages include codes which provide force fields (CHARMM, AMBER), those which do quantum chemistry calculations (NWCHEM, for example), those which organize molecular dynamics calculation for cluster computing (e.g. NAMD) and those which do image analysis (HELIX-FINDER for Cryo-EM data, for example). These packages are all research-group based and hence can both incorporate new ideas as they emerge in the community and remain usable by new scientists as they become trained in the field. There are organized community efforts to train new users, such as summer schools in computational methods in biophysics being run at various universities, for example. Alternative approaches to software development such as having a group of software developers work in relative isolation on a set of modules that a limited set of people have formulated at some fixed time-point is not appropriate in our view for a rapidly advancing, highly distributed yet organized, research community.

After repeated badgering of our briefers and after repeated attempts to look through the computational molecular biophysics literature, no truly compelling case emerged for HPC as deployed, for example, by the NNSA ASC program. The difficulties are the mismatch between scales at which we can be reasonably confident of the fundamental interactions (here atoms and electrons, at scales of angstroms and femtoseconds) and scales at which we want to understand biomolecular structure and function (tens to hundreds of nanometers, milliseconds and longer). This means that large scale ab initio simulations are most likely not going to dominate the field and that it will be difficult for massive capability platforms to make a huge difference.

Instead, we recommend vigorously supporting research in this area with something like the current mix of computation resources. There needs to be a continuing investment in algorithms and in machine architecture issues so that we can overcome the "serial bottleneck" and can seamlessly accomplish multi-scale modeling, as informed by the scientific need. The digital ribosome is not feasible today as a computation grand challenge, but is sufficiently close to deserve further scrutiny as our understanding improves.

4 GENOMICS

In this section we provide some perspectives on the role of HPC in genomics. We conclude this section with an assessment of a potential grand challenge that connects developments in genome sequencing with phylogenetic analysis: determination of the genome of the last common ancestor of placental mammals.

4.1 Sequence Data Collection

Presently, raw DNA sequence information is deposited in an international trace archive database managed by US National Center for Biotechnology Information. Each "trace" or "read" represents about 500-800 bases of DNA [31]. Most reads are produced by "shotgun" sequencing, in which the genome of interest is randomly fragmented into pieces of a few thousand bases each, and the DNA sequence at the ends of these pieces is read. The Joint Genome Institute (JGI) at DOE is one of the top four producers of DNA reads in the world. The other three are NIH funded labs. JGI contributed roughly 20 million DNA traces in the three months ending July 2004, which is about 25% of the worldwide production that quarter. The cumulative total JGI contribution to the trace archive as of July 2004 was approximately 46 million traces, representing about 10% of total worldwide contributions. Approximately 80% of the DNA in the trace archive was generated by the top four labs.

Beyond its great biomedical importance, extensive DNA sequencing has the potential to give us significantly greater depth in understanding the biodiversity on this planet and how it has evolved. In addition to sequencing the (nearly) complete genomes of hundreds of individual species, the shotgun sequencing methods have been applied to the analysis of environment samples, where genome fragments from a complex mixture of species living in a given ecosystem are all obtained at once from a single experiment [42, 41]. It is anticipated that in the near future these methods will generate significant amounts of genome data from organisms very broadly distributed over the tree of life. Data from environmental sequencing efforts could be used to identify new species and new members of gene families, with potential applications in medicine, ecology and other areas.

Venter et al. [42] report obtaining more then 1 million new protein sequences from at least 1800 prokaryotic species in a single sample from the Sargasso Sea. The method is remarkably successful for species that are abundant in the sample and exhibit little polymorphism, i.e. DNA differences between individuals.

The polymorphism issue is an important one. In the Sargasso Sea study, some species had as little as 1 single nucleotide polymorphism (SNP) in 10,000 bases. A length-weighted average of 3.6 SNPs per 1000 bases was obtained for all species from which they could assemble genomic DNA into large contiguous regions ("contigs"). A relatively low SNP rate such as this is necessary if one is to reliably assemble individual reads into larger contigs without crossing species or getting confused by near duplicated sequence within a single species. Larger contigs are useful for many types of analysis. It is unclear how many species are not analyzable in an environmental sample of this type because of high polymorphism rates. Polymorphism rates as high as 5 SNPs per 100 bases can occur in neutrally evolving sites in eukaryotes such as the sea urchin (Eric Davidson, personal communication). Such a high rate of polymorphism makes it difficult to correctly assemble contigs across neutral regions even in a pure diploid sample from a single eukaryotic individual. The situation is much worse in an environmental sample. Still, there is some hope of assembling somewhat larger contigs in regions that are protein coding or produce structural RNA if strong purifying selection within the species is constraining the DNA sufficiently (e.g. in ribosomal

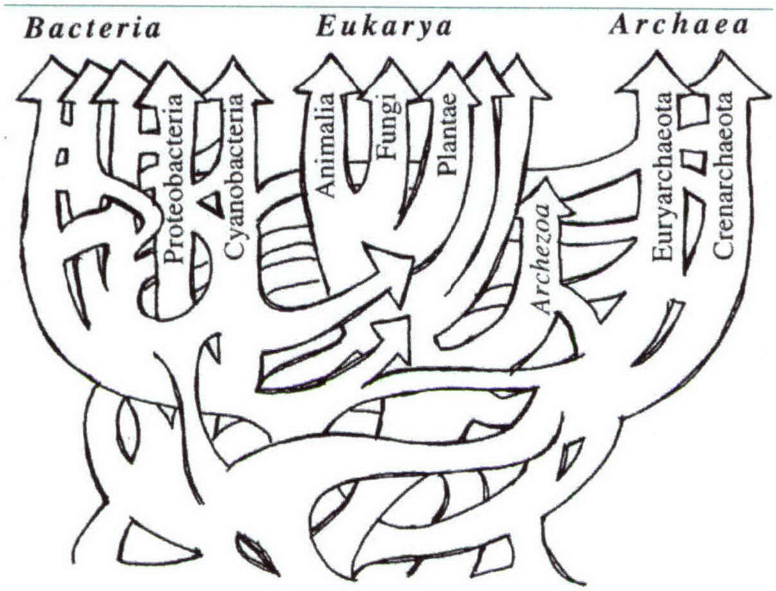

Figure 4-10: A depiction of the tree of life indicating the complications caused by reticulated evolution.

RNA genes, which are typically used to identify species). Because there will nearly always be some neutral polymorphic sites intermingled with the constrained sites, better, more "protein-aware" or "RNA-aware" methods of sequence assembly will be needed to fully exploit environmental sequence data by producing the largest possible contigs.

There is significant synergy with the DOE sequencing programs and the NSF "Tree of Life" initiative, whose goal is to catalog and sort out the phylogenetic relationships among the species present on our planet. This project is even harder than one might expect, because contrary to the original picture of Darwin, it is clear that species relationships are complicated by reticulated evolution, in which DNA is passed horizontally between species, creating a phlyogenetic network instead of a simple tree (see Figure 4-10). While rare in animals, this is especially prevalent in the bacterial kingdom, an area where DOE has significant opportunity in light of NIH's focus on metazoan genomes and NSF's focus on plant genomes. Significant sequencing of bacterial genomes is needed to sort this issue out. Simple analysis based on sequencing of a few common functional elements from each species' genome, such as the ribosomal RNA genes, will not suffice.

4.2 Computational Challenges

There are a number of computational challenges related to the efforts described in the previous section.

4.2.1 DNA read overlap recognition and genome assembly

As discussed above, individual DNA reads must be assembled into larger genome regions by recognizing overlaps and utilizing various kinds of additional constraints. This has been challenging even for DNA reads from a single species. In environmental sequencing, this must be done without mixing DNA from different species. As mentioned above, sparse sampling of DNA from many species in the more complex environmental samples, coupled with high rates of polymorphism within specific species presents a significant obstacle here.

4.2.2 Phylogenetic tree reconstruction

There have been potentially significant algorithmic advances for reconstructing phylogenetic trees, including meta-methods for improving the performance of current algorithms. But the data sets on which this development can take place are still limited and there does not yet seem to be sufficient understanding the nature of real world problems to create useful synthetic data sets. The current assessment is that reconstructing large phylogenetic trees will require potentially large parallel machines at some point in the future and further the more efficient algorithms may require more conventional supercomputer architectures. One should monitor the developments here closely over the next two or three years. More specific challenges include finding improved techniques for the major hard optimization problems (maximum parsimony and maximum likelihood) in conventional phylogenetic inference, as well as dealing with higher level analysis of whole genome evolution, with insertions, deletions, duplications, rearrangements and horizontal transfer of DNA segments.

4.2.3 Cross-species genome comparisons

Orthologous genomic regions from different species must be detected and aligned in order to fully identify functional genomic elements (protein-coding exons, non-coding RNA sequences, and regulatory sequences) and to study their evolution from a common ancestor. In evolutionarily close species, e.g. for the human and mouse genomes, genomic alignment and comparison can be done solely at the DNA level, although further analysis of the robustness of these alignments is warranted. As an example of the computational capacity required to do this, running on the 1000 CPU commodity hardware cluster at David Haussler's laboratory at UCSC, it takes Webb Miller's BLASTZ program 5 hours to compare and align the human and mouse genomes. Note that the requirements here are for capacity. Typically, these computations are "embarrassingly parallel".

In more distant species comparisons, e.g. human to fly, too much noise has been introduced by DNA changes to reliably recognize orthologous DNA segments by direct matching of DNA sequences. In this case it is common to first identify the protein coding regions in each species' DNA and then compare these as amino acid sequences, which exhibit many fewer changes than does the underlying DNA due to the redundancy of the genetic code. In principle, these protein level alignments could be projected back onto the

DNA sequences and even extended some (perhaps short) distance into the nearby non-coding DNA. This would be a useful algorithmic and software development. Production of alignments anchored off conserved non-coding elements, such as non-coding RNA genes would also be of great value. This presents a significant computational challenge and depends greatly on obtaining a better understanding of the molecular evolution of some of the various classes of functional non-coding genomic elements. Finally, in species with introns, which includes virtually all multicellular organisms, the identification of protein coding genes is significantly more complicated, and it appears that combined methods of comparative alignment and exon detection are needed [27]. an area of active research. At present, code developed in Haussler's lab using phylogenetic extensions of hidden Markov models is used to identify likely protein coding regions. It takes days to run on the human, mouse and rat genomes on their 1000 CPU cluster. Again, the challenge here is to deploy sufficient capacity.

4.2.4 Data Integration

To give genome sequences maximum utility, other types of biomolecular data must be mapped onto them, and made available in a common database. These types of data include cDNA sequences (a more direct window into the RNA sequences made by the species), gene expression levels under various conditions, evidence of protein-DNA interactions at specific sites (e.g. ChIP-chip data), etc. Web-based, interactive distribution of these data provides an opportunity to reach a large research audience, including labs less proficient in software development. This need for data federation and searchability appears in several other contexts in this report.

4.3 A Potential Grand Challenge - Ur-Shrew

An example of a grand challenge in computational genomics would be the reconstruction of the genome of the common ancestor of most placental mammals, a shrew-like animal that lived more than 75 million years ago. The idea would be to infer the DNA sequence of this ancestral species from the genomes of living mammals. This challenge involves a number of the areas mentioned above, including genome sequence assembly, whole genome sequence alignment and comparison, and inference of phylogenetic relationships from sequence, as well as areas not discussed, such as the detailed inference of specific molecular changes in the course of evolution. Recent work by Blanchette, Miller, Green and Haussler has indicated that with complete genomes for 20 well-chosen living placental mammals, it is likely that at least 90% of an ancestral placental genome could be computationally reconstructed with 98% accuracy at the DNA level [5]. Combined with the identification of the functional elements in mammalian genomes, including the protein-coding genes, RNA genes, and regulatory sequences, a reconstructed ancestral genome would provide a powerful platform for the study of mammalian evolution. In particular, it would allow us to identify the core molecular features that are common to and conserved in placental mammals, as well as the features have evolved to define the separate lineages, including the human lineage.

There are between 4000 and 5000 species of mammals currently identified, with the exact number still being the subject of debate. Mammals are not the most speciose animal even among vertebrates, where several groups have greater species counts according to present estimates; reptiles (~ 7000 species), birds ($\sim 10^4$ species) and fishes ($\sim 2.5 \cdot 10^4$ species). Of course numbers for various groups of invertebrates are much larger, such as molluscs ($\sim 8 \times 10^4$ species) and insects ($\sim 10^6$ species). The more living descendant species that are available, the more accurately one can reconstruct the ances-

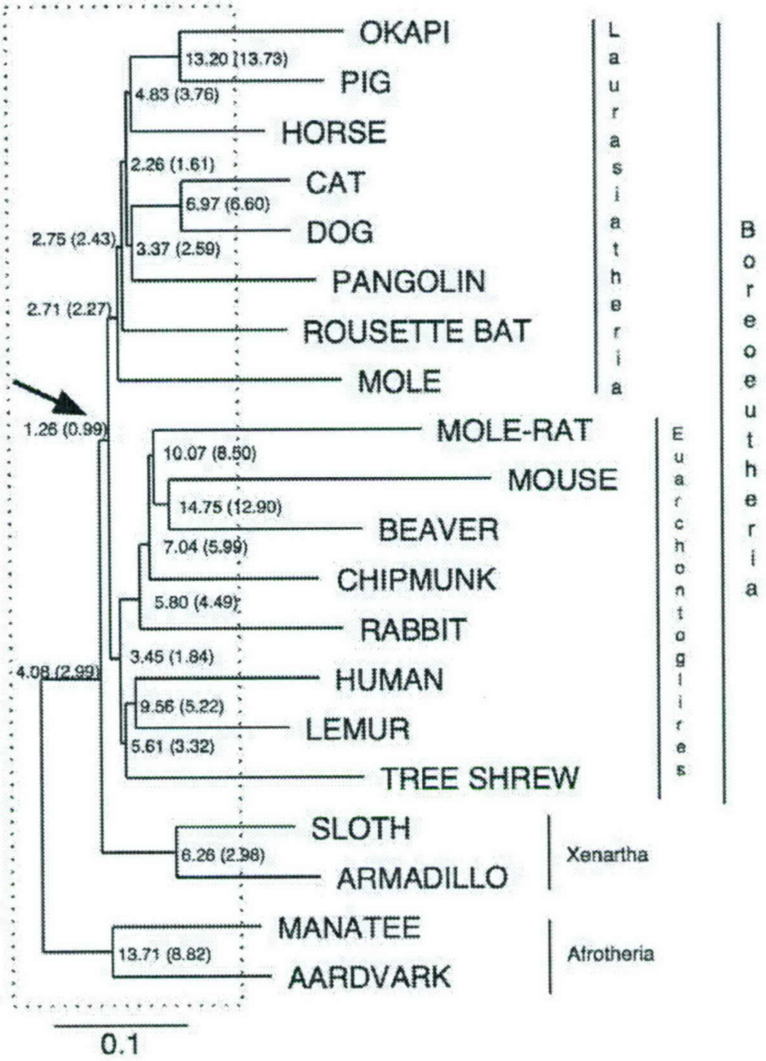

Figure 4-11: Base-level error rates in reconstruction of DNA from different placental ancestors. These are estimated from simulations in [5]. The numbers in parentheses are fraction of incorrect bases not counting repetitive DNA. Scale of branch lengths is in expected base substitutions per site. The arrow indicates the Boreoeutherian ancestor.

tral genome. However, the number of living species is not the only relevant parameter in determining how accurately one can reconstruct an ancestral genome. The time (or more specifically, the time multiplied by evolutionary rate) back to the common ancestor is very important, as is the topology of the phylogenetic tree. Better reconstructions are usually obtainable for collections whose last common ancestor existed at a time just before a period of rapid species diversification [5]. The rapid radiation of placental mammals (3800 species), right after the extinction at the Cretaceous-Tertiary boundary approximately 65 million years ago, makes a placental ancestor an attractive candidate. The target ancestor would be one that lived some time before this event, e.g. at the primate-rodent split, estimated at 70 million years ago [12], or earlier. One attractive choice is the boreoeutherian ancestor [5], a common ancestor of a clade that includes primates, rodents, artidactyls (including, e.g. cows, sheep, whales and dolphins), carnivores and other groups, which may have lived up to 100 million years ago (see Figure 4-11). In contrast, the last common ancestor of all mammals, including marsupials and monotremes, is thought to date back to the Triassic Period (195-225 million years) [12].

The Cretaceous-Tertiary extinction event is estimated to have killed about 50% of all species. However, it was not as severe as the Permian-Triassic extinction event of 252 million years ago, during which about 95% of all marine species and 70% of all land species became extinct. This is considered to be worst mass extinction on Earth so far. It would be an even greater challenge to attempt reconstruction of an ancestral genome from this time, but the magnitude of DNA change since this time is likely to be such that much necessary information will have been irrevocably lost.

To test the accuracy of a reconstructed genome, it would desirable to obtain actual DNA samples from ancestral species, hopefully from most major subclades and ideally from the most ancient ancestors possible. There have been claims made that DNA may be found in preserved ancient bacteria or even in dinosaur bones, but these claims remain highly controversial at best. The pre-fossil forests of Axel Heiberg Island in the Canadian Arctic yield mummified samples of bark and wood from trees which date back over 48 million years. The samples are organic. The unusual environmental history that created these samples could well have created samples of organic matter in similar stages of preservation from other organisms. However, whether any useful DNA sequence data can be obtained from these remains open. On the other hand, there is a credible claim by a team of Danish scientists for plant and animal DNA dating between 300,000 and 400,000 years ago, obtained from drilling cores collected in Siberia. However, others have argued that no reliable DNA can be obtained from remains more than 50-100 thousand years old [4, 25]. Given that the most recent evolutionary branch point with a living species related to humans, namely the chimpanzee, occurred more than 5 million years ago, this means that options for testing the accuracy of the computationally reconstructed genome sequence of a species ancestral to us by recovering a sample of that or a closely related species' DNA are limited.

Another approach to experimentally validating the ancestral sequence would be to synthesize individual genes from it, clone them into a mouse model, and test their activity *in vivo*. This will require advances in DNA synthesis technology, but is not out of the question. However, such a test could never prove that the reconstructed gene was correct, only that it is functional. Further, there may be problems due to the fact that the other genes, including those that have close interactions with the reconstructed ancestral gene, would still be murine genes. Nevertheless, the information gained from such tests would be useful.

Our conclusion is that the "Ur-Shrew" grand challenge may be one that is worthwhile and could be pursued quite soon. Assuming that NIH's plans to sequence a broad sampling of placental mammals are carried out, and the estimates from [5] hold up, the data required to get a reasonably useful reconstructed ancestral placental mammalian genome will soon be available. The most pressing need will then be for more powerful computational comparative genomics and phylogenetic analysis methods, as discussed in the sections above. The HPC requirements for this project seem to be for increased computational capacity, not computational capability. In other words, if this project, or a related project with species sequenced by DOE were to be undertaken, DOE should encourage the acquisition and use of commodity clusters, either by individual labs or as part of a national facility. This holds for many other challenges one might consider in the areas of genomics as well.

5 NEUROSCIENCE

5.1 Introduction

The field of neuroscience encompasses a large number of scales of interest. This is illustrated in Figure 5-12 as described to us by briefer T. Sejnowski. The figure displays a length scale hierarchy starting at the most basic level with the molecules that form neural synapses. At the next level of organization are neurons. One can then formulate higher levels of organization composed of networks of neurons which then map to regions of the brain. Ultimately, the goal is to understand how all these interacting scales come together to dictate the behavior of the central nervous system. Contributions to computational neurobiology occur at every level of this hierarchy. Given the breadth of the area, it is impossible to cover throughly the field in this report. Instead, we describe here briefly several aspects of computational neuroscience as briefed to us by Mayank Mehta, Terrence Sejnowski, and Garret Kenyon. Each of these briefings raise important issues relative to requirements for high performance computation. We close this section with a discussion of a potential grand challenge in computational neuroscience that attempts to model the retina.

A central issue raised in the briefing of Terry Sejnowski is the understanding of the mechanisms by which signaling takes place within the synapses connecting neurons. Neurons communicate through firing events between synapses. These firing events represent the release of various chemical transmitters which then activate a target neuron. The transmitters can also dynamically alter the operating characteristics of the signaling machinery itself. It is through this dynamic mechanism that various brain functions such as memory are accomplished. For example, in the formation of new

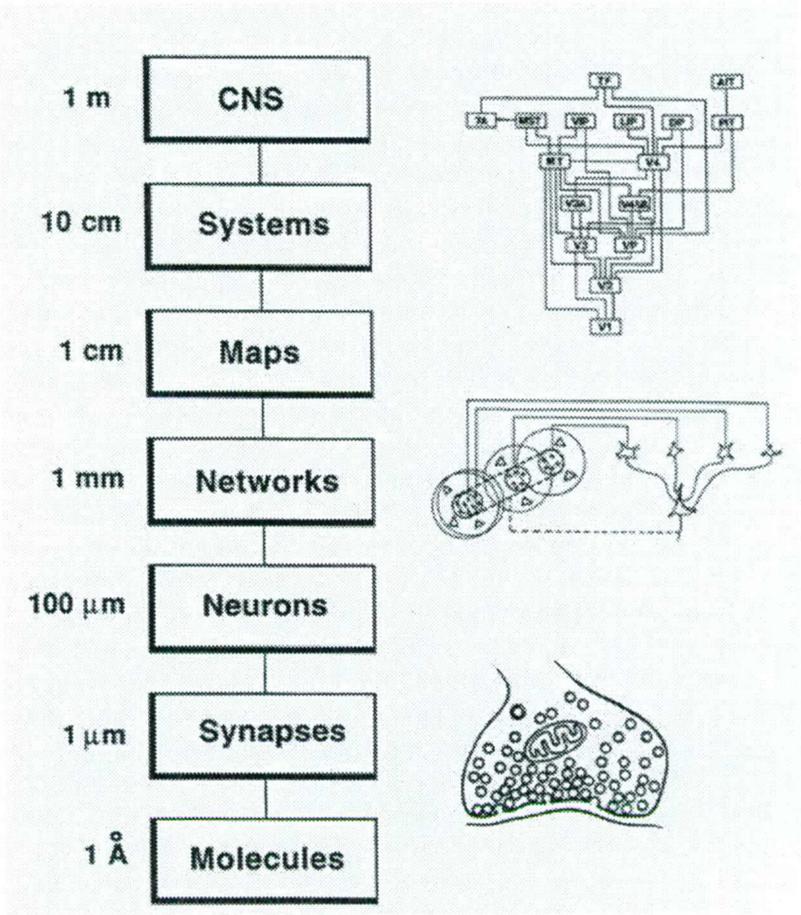

Figure 5-12: The neural hierarchy (from the briefing of Dr. T. Sejnowski.

memories it is thought that various synapses among the associated neurons are strengthened through this dynamic mechanism so as to encode the new memory for later retrieval. This dynamic updating of synaptic strength is referred to as "synaptic plasticity". Sejnowski described in his briefing recent work by Mary Kennedy and her coworkers on a complex of signaling proteins called the post-synaptic density which is located underneath excitatory receptors in the central nervous system. Kennedy's group has used a variety of techniques to elucidate the structure of these proteins and is now examining the interaction among these proteins in controlling transmission in the synapse and in effecting the phenomenon of plasticity. Some of the identified proteins are shown in figure 5-13. Sejnowski argues that such com-

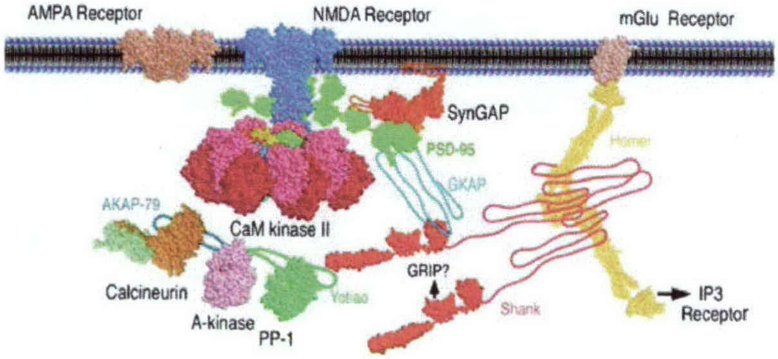

Figure 5-13: Signaling proteins in the post synaptic density. The figure is taken from work of Prof. Mary Kennedy as briefed to us by Prof. T. Sejnowski.

plex dynamics cannot be adequately modeled via a classical reaction-diffusion based model of the reaction dynamics. Instead it becomes necessary to take into account the complex geometry and the stochastic fluctuations of the various biochemical processes. In this approach, diffusion is modeled via a Monte Carlo approach applied to the individual molecules that participate in the biochemical reactions. Reactions are also treated stochastically using a binding rate. As the random walk proceeds, only molecules that are in close proximity will react and then only if the binding rates are favorable. The contention is that this flexibility in the ability to prescribe the general in-vivo geometry and the more detailed approach to the reaction dynamics is essential to properly describing the reaction dynamics. Kennedy and her group are able to provide estimates to the Sejnowski group of the average numbers of each molecule that is present as well as anatomical data of various neurons in portions of the brain. The computational requirements here certainly require high performance computation and the Sejnowski group has developed the MCell program as a tool to numerically perform the required stochastic simulation in a prescribed geometry of interest. There is great value in such studies as they can either point the way to obtaining better "continuum models" of plasticity or can help in the development of more sophisticated synaptic modeling strategies. It should be pointed out, however,

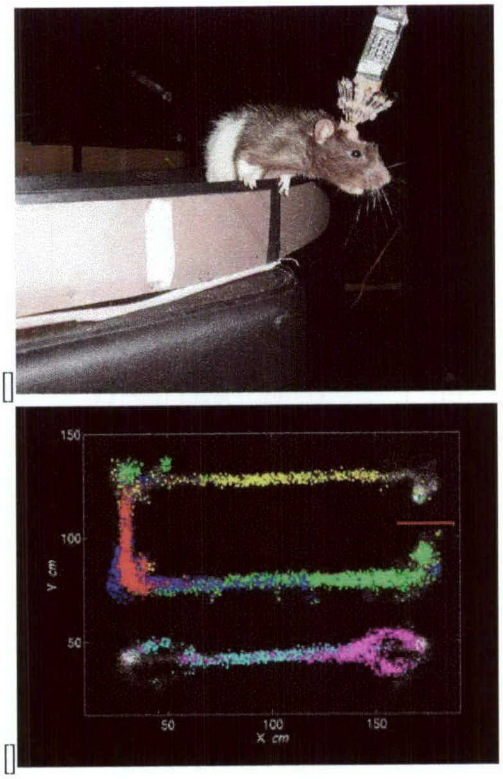

Figure 5-14: (a) Tetrode configuration to study multi-neuron measurements in rats. (b) Activation of various neurons as a function of the location of the rat.

that this simulation is at the subcellular level and so the path to integrating this detailed knowledge to the cellular level (or even beyond to the network level) is unclear at present. Thus, while HPC is clearly helpful here, we do not see that this approach could be the basis for large scale simulation of neural processing which is presumably the ultimate goal. As in the case of protein folding, some sort of "mesoscopic" approach must be developed (possibly with the assistance of tools like MCell). If such an approach can be developed, then large scale computation of neural networks informed by such modeling becomes possible and at this point a large investment in HPC capability may well be required, at the present time, however, we see this area as being better served by deployment of capacity platforms so that a number of simulation approaches can be investigated.

The phenomenology of synaptic plasticity can also be explored experimentally and in this regard we were briefed by Prof. Mahank Mehta who described the use of multi-neuron measurements by simultaneous recording of EEG signals from over 100 neurons in freely-behaving rats over a period of several months using tetrodes. An example of this approach is shown in Figure 5-14. The benefit of this approach is that is it possible to understand correlations among neurons as learning occurs. Mehta's results show that the activity of various hippocampal neurons depend on the rat's spatial location, that is, that the rat hippocampus apparently has "place cells" to help it reason about its spatial location. The main implication for our study of HPC is that such measurements require the ability to store, manipulate and ultimately to reason about an enormous amount of data. The neurophysics community has understood this and a number of Grid-based projects have been initiated.

5.2 A Potential Grand Challenge – The Digital Retina

In this section we will consider the case for large scale simulation of the retina as a possible grand challenge in the area of neuroscience. As we will argue, the retina in primates and other animals meets the criteria for a grand challenge quite well. As noted in the overview, to qualify for our category of grand challenge a problem should have the following features:

- A one decade time scale
- Grand challenges cannot be open-ended
- One must be able to see one's way, albeit murkily, to a solution.
- Grand challenges must be expected to leave an important legacy.

We begin by considering our understanding of the current state of knowl-

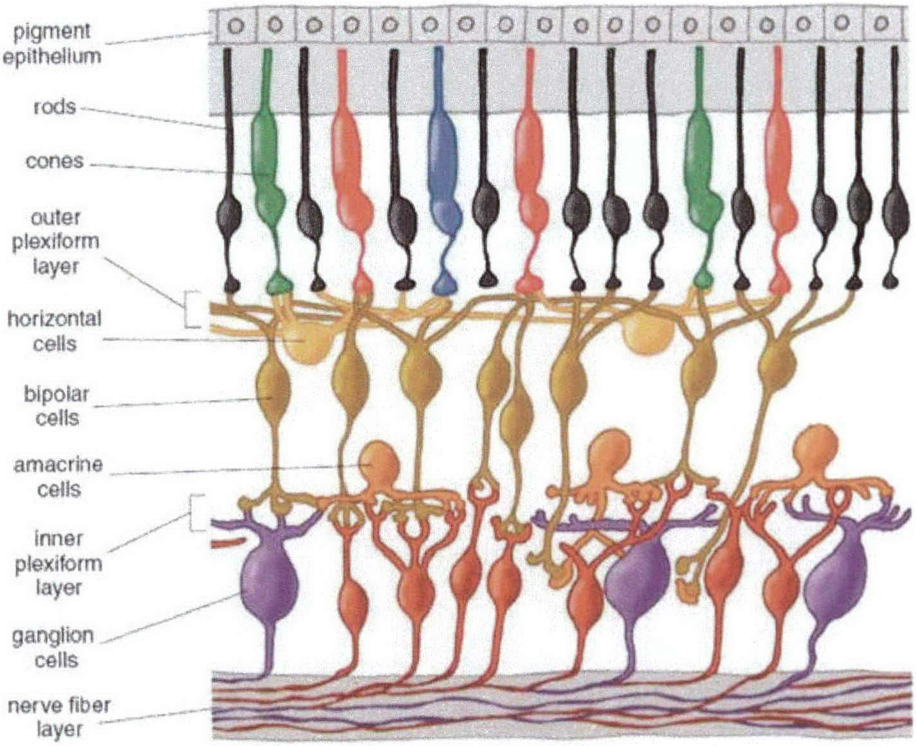

Figure 5-15: Architecture of the retina from [24]. Cells in the retina are arrayed in discrete layers. The photoreceptors are at the top of this rendering, close to the pigment epithelium. The bodies of horizontal cells and bipolar cells compose the inner nuclear layer. Amacrine cells lie close to ganglion cells near the surface of the retina. Axon-to-dendrite neural connections make up the plexiform layers separating rows of cell bodies.

edge as regards the retina. Our assessment is that the state of understanding is rather well advanced. As explained article of Kolb [24], many of the detailed cellular structures in the retina are well established. In the figure from Kolb's article (Figure 5-15) we see the layered structure of the retina which takes light input to the rods (senses black and white) and cones (senses red, green, and blue in primates) and through modulation through the bipolar cells and ganglion cells transforms the input to spike trains propagated to the brain along the optic nerve. There are roughly 130 million receptors and 1 million optic nerve fibers. Kolb notes that we can say we are halfway to the goal of understanding the neural interplay between all the nerve cells in the retina. We interpret this as meaning that experimental scientists are well along in collecting, collating, and fusing data about the structure and interaction of the neurons in the retina.

The next step in our outline is also reasonably well established. There seems to be general agreement about the structure of many retinas in various animals, and there is the beginning of a web based documentation on retinal genes and disorders: (see for example [21]). We could not find a database of neural structures in various animals along with details about the electrophysiology of the neurons in the circuits. So, this step in the development of useful models requires further development. This aspect of the grand challenge certainly does not need high performance computing. A database in this arena could be assembled consisting of experimentally observed spike trains propagating along optical nerve fibers associated with some class of agreed-upon test scenes presented to experimental animals.

We next address the issue of simulation. Here, one can find many models for a few photoreceptors and associated bipolar, horizontal, amacrine, and ganglion cells, and even excellent work building small pieces of the retina in silicon. The paper in [9] is a recent example of this. We have not found any really large scale model of the retina in software or in hardware. If one wishes to simulate the whole human retina with 125 million receptors and associated processing neural circuitry leading to 1 million nerve fibers carrying spike visual information trains down the optical fiber, then to represent one minute of retinal activity with one hour of computing time one will need approximately 7-10 TFlops. This resolves the behavior of realistic neurons at a temporal resolution of 50 microseconds. The problem is eminently parallelizable as the computing units in the retina are similar in structure, not in actual physical density. Equivalently, for model development and parameter exploration, one could use the same computational power for a second of retinal activity realized in one minute. This level of computing power is commercially available today in a 1024 (dual processor) node IBM e1350.

This is somewhat beyond conventional clusters found in many laboratories, but requires no specialized computer development. Delivery of a 128 node 1.1 TFlop system was taken recently by NCAR and performance at the level of 7-10 TFlops has been achieved by several groups including the national laboratories several years ago. At this stage there is nothing we can say about prediction and design, though recent work (see [9]) may provide a start in this direction.

What is it that the DOE would need to do the develop the Retinal Decade (the 10 year period for the Digital Retina grand challenge)? The key ingredients go well beyond the required computational facility which would be achievable using present-day HPC resources. It would require an organization with extensive experimentation, as emphasized in the outline of a grand challenge in life sciences, that is well-coupled to the numerical modeling. The JGI is perhaps a model for this in that the sequencing machines were a critical part of the story but not the only critical part. The organization, training, well-defined goal setting, and a long term effort were critical.

It is appropriate to ask why one ought to consider the retina and not, for example, the entire visual system or, even, the cortex? The latter systems are simply not "ready for prime time" as a grand challenge in our view. Item one on the list of grand challenge criteria is drastically incomplete; the knowledge of the anatomy of the full visual system is reasonably known, though certainly not as well as the retina alone, and the detailed electrophysiology needed to make realistic models is absent. A similar situation holds for the cortex as a whole, though even there the anatomy is not fully developed.

The retina is a processing system which dissects a visual scene and transforms it into spike trains propagated along the optic nerve. If we can understand, in simulation and predictive models, how this is done in detail and through perturbations on that model why it is done the way nature does it and what other variations on this theme might exist, we will have for the first time fully characterized a neural circuit more complex than collections of tens to perhaps a few thousand neurons in invertebrate systems.

Further, we will have provided the basis for design principles for other visual processing systems using the ingredients of the model system. Our ability to go from the modeling, and reverse engineering of the retina, to designing new systems using the principles discovered, would constitute an important understanding of a critical circuit in ourselves. This would surely have implications for treatment of disease which we do not attempt to draw here. In addition, it would have equally important uses in the design of optical sensing systems for robots useful in commercial and military environments.

6 SYSTEMS BIOLOGY

In this section we review some of the work that was briefed to us on systems biology. This is an important area now in its developmental stages. Although systems biology means different things to different people, most would agree that it concerns the functioning of systems with *many* components. Practitioners of systems biology today are working primarily on subcellular and cellular systems (as opposed to, say, ecological systems, which are in themselves also very interesting biologically as well as from a systems perspective). Articulated goals of this field include elucidating specific signal transduction pathways and genetic circuits in the short term, and mapping out a proposed circuit/wiring diagram of the cell in the longer term. The essential idea is to provide a systematic understanding and modeling capability for events depicted in Figure 6-16: when a cell interacts with some agent such as a growth factor or a nutrient gradient, a complex series of signaling events take place that ultimately lead to changes in gene expression which in turn results in the cellular response to the stimulus. An example may be the motion of a cellular flagellum as the cell adjusts its position in response to the gradient.

The information leading to the reconstruction of the wiring diagram that describes the cellular response programs includes

1. data from various high throughput technologies (e.g., DNA microarray, CHiP-on-chip, proteomics),
2. results from the vast literature of traditional experimental studies,
3. homology to related circuits/networks worked out for different organisms.

The desired output of these approaches is a quantitative, predictive compu-

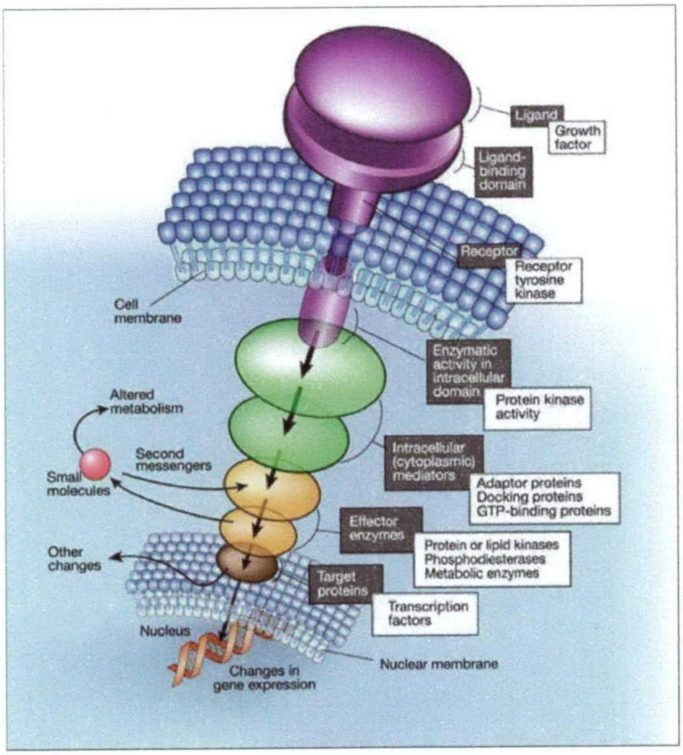

Figure 6-16: Cellular signaling - figure from presentation of Prof. Subramanian (UCSD).

tational models connecting properties of molecular components to cellular behaviors. Given this scope, a large part of systems biology being practiced today is centered on how to integrate the vast amount of the heterogeneous input data to make computational models. We were briefed by Prof. Shankar Subramanian who described the work of the Alliance for Cellular Signaling. This program aims to determine quantitative relationships between inputs and outputs in cellular behavior that vary temporally and spatially. The ultimate goal of this program is to understand how cells interpret signals in a context-dependent manner. One very important aspect is organizing the vast amount of data that arise in investigations of cellular signaling phenomena. As we comment later, quantifying the function and topology of cellular signaling networks is challenging. In order to assist with this goal, the Alliance has organized an enormous amount of data that can then be used by the community to test hypotheses on network structure and function. The computational requirements here are dictated mainly by the need to store and interrogate the data. We anticipate that over time there will be a need to make this type of data centrally available to researchers so that it can be easily obtained and assessed. This argues for a network-based information infrastructure linking searchable databases. In our briefings we heard several times about the need for such a facility - a "bioGoogle". Such a facility would be a significant undertaking and would certainly require multi-agency cooperation.

Other software development efforts include the M-Cell project (briefed to us by Dr. Terry Sejnowski) which focuses on modeling of neural synapses and the Biospice program as briefed to us by Dr. Sri Kumar of DARPA/IPTO. The goal of BioSpice is to provide a software platform to explore network dynamics as inferred from high throughput gene expression data. The major computational needs in these endeavors are

- bioinformatic processing of the high throughput data
- detailed stochastic simulation of network dynamics

There is little question that significant HPC requirements emerge in this endeavor even for bacterial systems such as E. Coli. Experiments indicate that, as the cell responds to a stimulus, the interconnection networks can become quite complex leading to complex optimization problems as one attempts to infer the network topology and system parameters from the data. If one then couples a predictive network model with a spatially and temporally realistic model of a cellular organism this will easily require HPC resources. Extrapolating in this way, the simulation requirements for multicellular organisms are even more daunting.

This would imply a ready arena for significant investment in HPC. It is, however, worthwhile to question the premise on which much of the above-mentioned program on systems biology is built upon. That is, that circuits and networks are, in fact, appropriate system-level descriptors that will enable quantitative, predictive modeling of biological systems. We discuss this in the section below and then close this section with discussion of a potential HPC grand challenge of simulating bacterial chemotaxis utilizing current approaches to systems biology.

6.1 The Validity of the Circuit Approach

To be sure, a network-based perspective beyond the single-gene paradigm of traditional molecular biology is crucial for understanding biology as a system. However, circuit diagrams are not necessarily the appropriate replacement. To appreciate this issue, it is instructive to examine the key ingredients that make circuit diagrams such a powerful descriptor for engineered electrical/electronic systems, e.g., integrated circuits:

- Components of an integrated circuit, e.g., transistors, are functionally simple. In digital circuits for example, a typical transistor (when properly biased) performs simple Boolean operations on one or two inputs. Moreover, the physical characteristics of a component relevant to its function can be summarized by a few numbers, e.g., the threshold voltage and gain. Thus, each component of a circuit can be quantitatively described by a standard model with a few parameters.

- These components operate in a well-insulated environment such that it is possible to specify only a few designated connections between the components; this property allows a clear definition of the connectivity of the component, i.e., the "circuit".

- Complexity of an integrated circuit arises from the iterated cascades of a large number of fast and similar components (e.g., 10^7 transistors switching at rates of typically a GHz). As the properties of the components are well characterized, the connectivity of these components is the principle determinant of system function.

- Even with the knowledge of a circuit diagram and the properties of the components, a complex circuit with various levels of feedback is still difficult to model quantitatively *ab initio* because circuits with cycles generally exhibit time-dependent behavior with unstable/unknown outputs. The proper function of a complex circuit generally requires its inputs to satisfy certain constraints. It is only with the knowledge of these constraints and the intended functions of the system can a complex circuit be understood and modeled quantitatively[4]

At present, it appears that few of the above features that make electronic circuits amenable to quantitative modeling are available today for evolved bio-molecular networks. We will illustrate the situation by examining the regulation of the *lac* operon [28], perhaps the best-characterized molecular control system in biology. The *lac* operon of *E. coli* encodes genes necessarily for the transport and metabolism of lactose, a carbon source which *E. coli* utilizes under the shortage of the default nutrient, glucose. The expression of the *lac* operon is under the control of the Plac promoter, whose apparent function is the activation of the operon in the presence of lactose, the "inducer". This is achieved molecularly via a double-negative logic as illustrated in Figure 6-17.

In the absence of the inducer, the transcription factor LacI binds strongly to Plac and prevents the access of the RNA polymerase required for transcription initiation. The inducer binds to LacI and drastically reduces its affinity for the specific DNA sequences contained in Plac, thereby opening up the promoter for transcription. The positive effect of lactose on the expression of the *lac* operon can be easily detected by modern DNA microarray

[4] In this context, we were briefed by Prof. Shuki Bruck of Caltech on possible principles for design of reliable circuit function even in the presence of feedback cycles. This work is in an early state and is reflective of the need to understand better biological "circuitry".

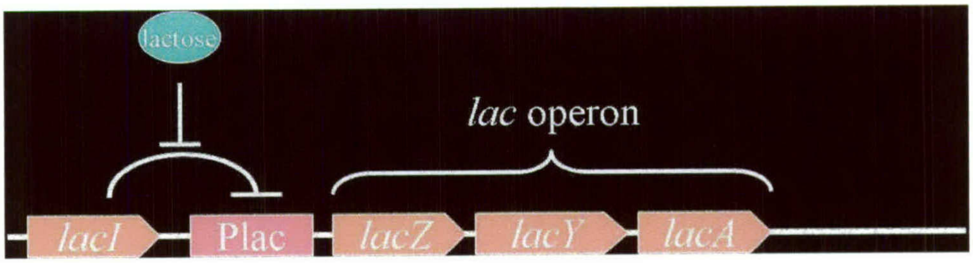

Figure 6-17: Schematic of the lac operon and its control by LacI and the inducer lactose.

experiments [47]. With some work, it is likely that the binding of LacI to Plac and its repressive effect on gene expression can also be discovered by high throughput approaches such as the ChIP-on-chip method [34]. Thus, the qualitative control scheme of Figure 6-17 is "discoverable" by bioinformatics analysis of high-throughput data. However, this information is far short of what is needed to understand the actual effect of lactose on the *lac* operon, nor is it sufficient to understand how the LacI-Plac system can be used in the context of large genetic circuits. We list below some of the key issues:

Difficulty in obtaining the relevant connectivity A key ingredient of the control of Plac by lactose is the fact that lactose cannot freely diffuse across the cell membrane. The influx of lactose requires the membrane protein lactose permease which is encoded by one of the genes in the *lac* operon [29]. Hence there is a positive feedback loop in the lactose-control circuit (cf. Figure 6-18). A small amount of lactose leaking into the cell due to a basal level of the lac permease will, in the presence of glucose shortage, turn on the *lac* operon which results in the infusion of more lactose. The positive feedback, coupled with a strongly nonlinear dependence of the promoter activity on intracellular lactose concentration, gives rise to a *bistable* behavior where individual cells switch abruptly between states with low and high promoter activities [32]. However, the onset of the abrupt transition is dependent on stochastic events at the transcriptional and translational level [43],

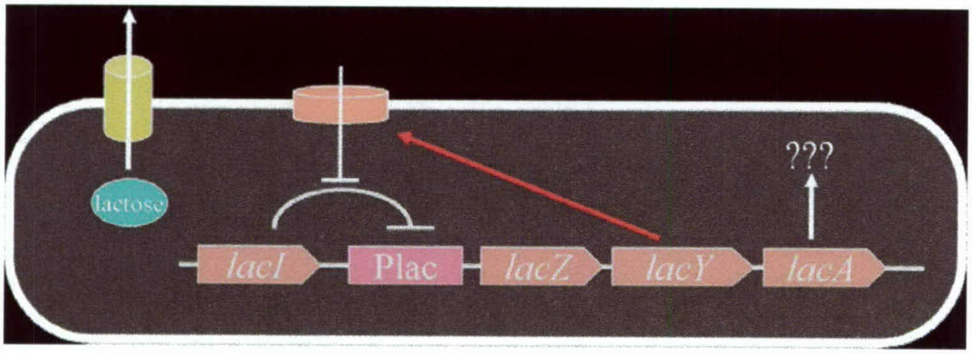

Figure 6-18: Influx of lactose requires the lac permease encoded by *lacY*.

so that at the population level, one finds instead a gradual increase of gene expression upon increase in extracellular lactose levels [22]. It is unclear how this positive feedback loop could have been determined by automated methods. It would require the knowledge of the intracellular lactose concentration and of the function(s) of the genes in the *lac* operon, which in turn require detailed biochemistry and genetics experiments. Without appreciating these issues, blindly fitting the smooth population-averaged behaviors to simple models of transcriptional initiation certainly will not generate reliable, predictive results. It should be noted that the function of the gene *lacA* in the *lac* operon is still not clear even today, and other mechanisms exist to change the intracellular lactose concentration (e.g., other diffusible inducers and the lactose efflux pump). Thus, further feedback control may well exist and the above circuit may still be incomplete.

Difficulty in reliable quantitation There are also problems with the characterization of the Plac promoter independent of the lactose transport problem. The gratuitous inducer isopropyl-b-D-thiogalactopyranoside (IPTG) can diffuse freely across cell membrane and bind to LacI, thereby activating transcription. The IPTG dependence of Plac activity has been studied by many groups. However, the results vary widely. For instance, reported values of fold-activation between no IPTG and 1mM IPTG can range from several tens to several thousands (see e.g., [37, 32, 47, 30]) on the same wild-type strain, and even more varied outcomes are obtained for different strains, under different (glucose-free) growth media, for different inducers and reporters. Biologists are usually aware of these differences, and the quantitative fold-changes are typically not taken seriously except that the promoter is "strongly activated" by the inducer. Thus, the problem of quantitation is not simply a "cultural issue" - that is, that biologists are not sufficiently quantitative. Rather, it is the complexity of the system that often makes reliable quantitation difficult. Also illustrated in this example is the danger of extracting quantitative results using automated literature search tools. Given the sensitive dependence of the systems on the details of the experiments, it is crucial to obtain the *precise context* of an experiment.

Difficulty in predicting function of a given circuit While dissecting real gene circuits *in vivo* is complicated by all sorts of unknown interactions, it is possible to set up artificial gene circuits and study their properties *in vivo* [19]. Given that the synthetic systems are constructed with reasonably well-characterized components which have clearly designated connections, they become a natural testing ground for quantitative computational modeling. A successful experiment in synthetic biology typically begins with a theoretically motivated circuit topology. It then takes several rounds of tweaking to make the construct behave in the designed manner. This is of course a standard practice for engineering of any man-made systems. However, the process also underscores how the behavior of the system depends on details, such that circuit topology is not a sufficient determinant of system properties. An explicit illustration of how the same circuit topology can give rise to different system-level behaviors is the experiment of [18] examining circuits consisting of the same 3 repressors but connected in a variety of different ways. They looked for the ability of these circuits to perform Boolean operations on two input variables (the concentrations of two ligands IPTG and aTc). What they found was that the same

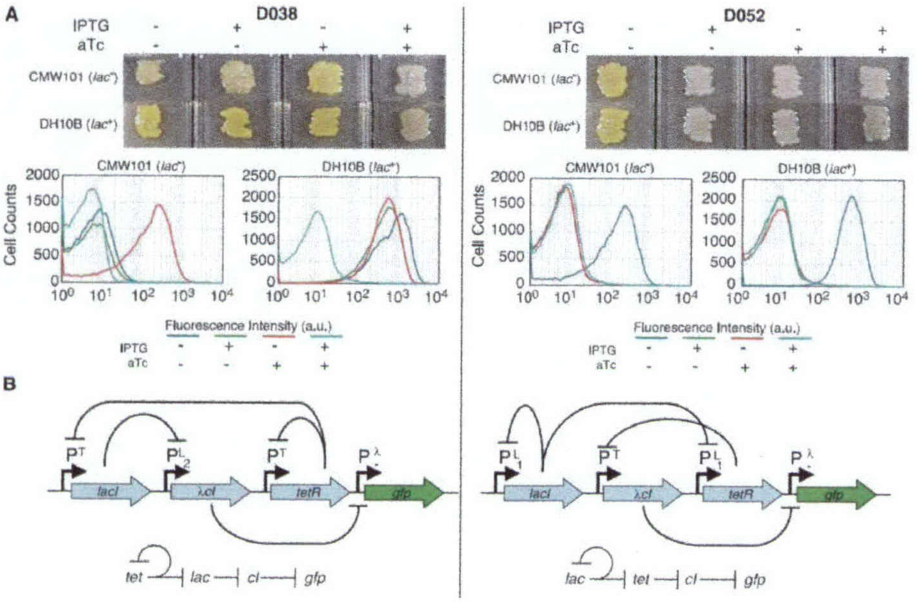

Figure 6-19: An explicit illustration of how the same circuit topology can give rise to different behaviors.

circuit topology can give rise to different logic functions (cf. Figure 6-19). In fact, out of the 15 distinct promoter combinations possible for their system, every circuit topology for which they made multiple realizations exhibited more than one type of behavior. Thus, the property of a circuit depended not only on its topology, but also other details that the circuit designers do not know about or over which they have no control. Possible factors include the relative strengths of expression and repression, leakiness of the different promoters, the turnover rates of the different mRNA and proteins, the order of genes on the plasmid, etc. Given that the promoters and genes used in the experiment (LacI, TetR, the lCI) are among the best characterized in molecular biology and yet naive expectations are not always realized, we believe it will generally be difficult to predict circuit properties based on connectivity information alone.

6.2 A Possible Grand Challenge: Bacterial Chemotaxis

With the above considerations we can consider the possible grand challenge of simulating a complex process such as bacterial chemotaxis. The problem has a well defined set of inputs, namely, the concentration field impinging on a cell membrane. The desired prediction is the dynamic response of the bacterial organism as a function of time. As discussed above, high throughput analysis has provided a wealth of data on the relevant molecular biology as the cell encounters various inputs in the medium.

However, as discussed above, the critical issue is a predictive approach to modeling cellular signaling. The cellular signaling process is at present not satisfactorily modeled, in our opinion, via a "parts list" connected via a discoverable network. The discussion of section 6.1 implies that additional investigation is clearly required into the details of the chemical networks that govern cellular signaling making investment of HPC resources to support a grand challenge in this area premature at the present time. There is no question that such a study is science-driven and its success would leave a clear legacy in the field. Indeed, once an appropriate modeling approach is identified that deals correctly with the issues identified on the previous section, a full spatially accurate model of the cell governed by an appropriate chemotaxis model would certainly require HPC resources in order to track the three dimensional response of the cellular system and its resulting dynamics in the medium.

7 CONCLUSIONS AND RECOMMENDATIONS

7.1 Introduction

In this section we provide a set of findings and conclusions for this study. We begin with some general observations and impressions about biology and the role of computation. First, biology is a data rich subject that poses challenges to the creation of models. For example, experiments turn up surprises all of the time and many zeroth order questions remain to be answered. This was underscored in many of our briefings (particularly those on systems biology). As a result, experiment remains the primary guide and information resource. From the (admittedly limited) set of briefings we received, we could not identify a situation in biology for which capability computation is currently a key factor limiting progress.

For computational modeling to be successful, there must be a plausible paradigm or model. For example, in particle physics, there is a long history of experimental and theoretical work leading up to universal agreement that a particular non-Abelian gauge-theory Lagrangian was a useful model to solve precisely and there was (and still is) extensive work to devise the proper numerical discretization. This work was essential for the productive application of large-scale computation. In almost all of the biology we heard about, the paradigm did not seem to be sufficiently firm to warrant large capability computational effort at this point.

Another principle is that the "right problem should be tackled at the right time with the right tools." As noted above, immature paradigms are a widespread feature at this point. But, in addition, supporting data are often lacking. For example, there is little doubt that neuronal interactions are the basis of mammalian brains, but the details of synaptic interactions, plasticity, etc. will be needed before large-scale modeling can be maximally productive. We do note that some special subsystems like the retina may be ready for large scale computation, but overall this fields remains driven by experiment and data collection. Similarly, metabolic pathways alone are not sufficient for systems-biology modeling; plausible values (or estimates) for reaction rates, diffusion constants, etc. will be necessary. At the present time, the right set of computational tools for the ongoing investigations appears to be at the level of workstations or clusters as opposed to capability platforms. We do note the potential importance of Grid computation.

We can generally identify a hierarchy of tasks to which computers and computation can be applied.

Data collection, collation, fusion Because biology is a data-rich subject with few mature paradigms, data are the touchstone for understanding. These data take many forms, from databases of sequence and structure to text literature. Further, the data are growing exponentially, due in part to advances in technology (sequencing capability, expression arrays, etc.) Collecting, organizing, fusing such data from multiple sources and making them easily accessible both to the bench researcher and the theoretician in a convenient format is an important and non-trivial information-science task, although not within the realm of traditional computational science.

Knowledge extraction The automated (or assisted) identification of patterns in large datasets is another large-scale computational task. Examples include genomic sequence homologies, structural motifs in proteins, and spike-train correlations in multi-electrode recordings. At some level, this activity must be guided by paradigms.

Simulation Here, a physical model is typically used to embody experimental knowledge. One obvious use to sufficiently encapsulate the existing phenomenological information. But more important is the understanding stemming from the construction and validation of the model.

Prediction With a validated model, one can make predictions. That is, what is the response of the system when it is changed or subject to external perturbation?

Design This is probably the highest level of computation. Here one investigates deliberate perturbations and/or combinations of existing systems to modify function. Validated models are essential at this level.

At present, our overall impression is that computation is playing an essential role in the first two aspects and increasing roles in the third. Given this emphasis, investments in capacity level and Grid-based computing seem most appropriate at this time. As modeling and understanding improve we expect to see much more utilization of computation to support simulation, prediction and ultimately, design.

7.2 Findings

Role of computation: Computation plays an increasingly important role in modern biology at all scales. High-performance computation is critical to progress in molecular biology and biochemistry. Combinatorial algorithms play a key role in the study of evolutionary dynamics. Database technology is critical to progress in bioinformatics and is particularly important to the future exchange of data among researchers. Finally, software frameworks such as BioSpice are important tools in the exchange of simulation models among research groups.

Requirements for capability: Capability is presently not a key limiting factor for any of the areas that were studied. In areas of molecular biology and biochemistry, which are inherently computationally intensive, it is not apparent that substantial investment will accomplish much more than an incremental improvement in our ability to simulate systems of biological relevance given the current state of algorithms and architecture. Other areas, such as systems biology will eventually be able to utilize capability computing, but the key issue there is our lack of understanding of more fundamental aspects, such as the details of cellular signaling processes.

Requirements for capacity: Our study did reveal a clear need for additional capacity. Many of the applications reviewed in this study (such as image analysis, genome sequencing, etc.) utilize algorithms that are essentially "embarrassingly parallel" algorithms and would profit simply from the increased throughput that could be provided by commodity cluster architecture as well as possible further developments in Grid technology.

Role of grand challenges: It is plausible (but not assured) that there exist suitable grand challenge problems (as defined in section 2.3) that will have significant impact on biology and that require high-performance capability computing.

Future challenges: For many of the areas examined in this study, significant research challenges must be overcome in order to maximize the potential of high-performance computation. Such challenges include overcoming the complexity barriers in current biological modeling and understanding the detailed dynamics of components of cellular signaling networks.

7.3 Recommendations

JASON recommends that DOE consider four general areas in its evaluation of potential future investment in high performance bio-computation:

1. Consider the use of grand challenge problems to make the case for present and future investment in HPC capability. While some illustrative examples have been considered in this report, such challenges should be formulated through direct engagement with (and prioritization by) the bioscience community in areas such as (but not limited to) molecular biology and biochemistry, computational genomics and proteomics, computational neural systems, and systems or synthetic biology. Such grand challenge problems can also be used as vehicles to guide investment in focused algorithmic and architectural research, both of which are essential to successful achievement of the grand challenge problems.

2. Investigate further investment in capacity computing. As stated above, a number of critical areas can benefit immediately from investments in capacity computing, as exemplified by today's cluster technology.

3. Investigate investment in development of a data federation infrastructure. Many of the "information intensive" endeavors reviewed here can be aided through the development and curation of datasets utilizing community adopted data standards. Such applications are ideally suited for Grid computing.

4. Most importantly, while it is not apparent that capability computing is, at present, a limiting factor for biology, we do not view this situation as static and, for this reason, it is important that the situation be revisited in approximately three years in order to reassess the potential for further investments in capability. Ideally these investments would be guided through the delineation of grand challenge problems as prioritized by the biological research community.

A APPENDIX: Briefers

Briefer	**Affiliation**	**Briefing title**
David Haussler	UC Santa Cruz	Genomes primer
Mayank Mehta	Brown University	Neurophysics of learning
Terry Sejnowski	Salk Institute	Modeling mesoscopic biology
John Doyle	Caltech	Systems biology
Garrett Kenyon	Los Alamos Nat'l Lab	Computational neuroscience
Mike Colvin	Livermore and UC Merced	Molecular dynamics
Eric Jakobsson	NIH	The BISTI Initiative
Shankar Subramanian	UCSD	Alliance for cell signaling
David Dixon	Univ. Alabama	Computational biochemistry
Wah Chiu	Baylor Univ.	Imaging and crystallography
Dan Rohksar	Lawrence Berkeley Lab	Sequencing of Ciona
Peter Wolynes	UCSD	Protein folding
Steve Mayo	Caltech	Protein structure and design
Jehoshua Bruck	Caltech	Biological circuits
John Wooley	UCSD	Advanced computation for biology
Nathan Baker	Washington Univ.	Multiscale modeling of biological systems
Klaus Schulten	Univ. Illiois	Theoretical molecular biophysics
Sri Kumar	DARPA	DARPA Biocomputation
Tandy Warnow	Univ. Texas (Austin)	Assembling the tree of life

[13] R. Elber, A. Ghosh, A. Cardenas, and H. Stern. Bridging the gap between reaction pathways, long time dynamics and calculation of rates. *Adv. Chem. Phys.*, 126:93–129, 2003.

[14] M. J. Field, P. A. Bash, and M. Karplus. Combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.*, 11:700–733, 1990.

[15] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal. Quantum monte carlo simulations of solids. *Rev. Mod. Phys.*, 73:33–83, 2001.

[16] J. C. Grossman, M. Rohlfing, L. Mitas, S. G. Louie, and M. L. Cohen. High accuracy many-body calculational approaches for excitations in molecules. *Phys. Rev. Lett.*, 86:472, 2001.

[17] J. C. Grossman, E. Schwegler, E. W. Draeger, F. Gygi, and G. Galli. Towards an assessment of the accuracy of density functional theory for first principles simulations ofwater. *Chem. Phys.*, 120:300–311, 2004.

[18] C. Guet et al. Combinatorial synthesis of genetic networks. *Science*, 296:1466–1470, 2002.

[19] J Hasty, D McMillen, and JJ Collins. Engineered gene circuits. *Nature*, 420:224–230, 2002.

[20] International Human Genome Sequencing Consortium (IHGSC). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[21] Retina International. http://www.retina-international.org/sci-news/anim-dat.htm.

[22] A Khlebnikov and JD Keasling. Effect of lacy expression on homogeneity of induction from the p(tac) and p(trc) promoters by natural and synthetic inducers. *Biotech Prog.*, 18:672–4, 2002.

[23] W. Kohn and L. J. Sham. Self consistent equations including exchange and correlation effects. *Phys. Rev. A*, 140:1133–1138, 1965.

[24] H. Kolb. How the retina works. *American Scientist*, 91:28–35, 2003.

[25] I. Marota and F. Rollo. Molecular paleontology. *Cell Mol Life Sci.*, 59:97, 2002.

[26] G. Monard and K. M. Merz Jr. Combined quantum mechanical/molecular mechanical methodologies applied to biomolecular systems. *Accts Chem. Res.*, 32:904–911, 1999.

[27] Brent MR and Guigo R. Recent advances in gene structure prediction. *Curr Opin Struct Biol.*, 14, 2004.

[28] B. Muller-Hill. The *lac* operon: A short history of a genetic paradigm. 1996.

[29] A Novick and M Weiner. Enzyme induction as an all-or-none phenomenon. *PNAS*, 43:553–566, 2001.

[30] S. Oehler, E. R. Eismann, H Kramer, and B Muller-Hill. The three operators of the lac operon cooperate in repression. *EMBO J*, 9:973–979, 1990.

[31] National Institutes of Health. http://www.ncbi.nlm.nih.gov/traces.

[32] E M Ozbudak, Thattai M, H Lim, BI Shraiman, and A van Oudenaarden. Multistability in the lactose utilization network of *Escherichia coli*. *Nature*, 427:737–740, 2004.

[33] VS Pande. http://folding.stanford.edu.

[34] B Ren, F Robert, JJ Wyrick, O Aparicio, EG Jennings, I Simon, J Zeitlinger, J Schreiber, N Hannett, E Kanin, TL Volkert, CJ Wilson, SP Bell, and RA Young. Genome-wide location and function of dna binding proteins. *Science*, 290:2306–2309, 2000.

[35] C. Rovira and M. Parinello. Harmonic and anharmonic dynamics of fe-co and fe-o_2 in heme models. *Biophys. J.*, 78:93–100, 2000.

[36] R. Schwitters et al. Requirements for ASCI. Technical report, JASON-MITRE, 2003.

[37] Y. Setty, A.E. Mayo, M.G. Surette, and U. Alon. Detailed map of a cis-regulatory input function. *Proc. Natl. Acad. Sci.*, 100:7702–7707, 2003.

[38] K. Siva and R. Elber. Ion permeation through the gramicidinchannel: atomically detailed modeling by the stochastic differenceequation. *Proteins Structure Function Genetics*, 50:63–80, 2003.

[39] C. Stubbs et al. The computational challenges of medical imaging. Technical report, JASON-MITRE, 2004.

[40] J. Tse. Ab initio molecular dynamics with density functional theory. *Annu. Rev. Phys. Chem.*, 53:249–290, 2002.

[41] J Tyson et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428:37–34, 2004.

[42] J C Venter et al. Environmental genome shotgun sequencing from the sargasso sea. *Science*, 304:66, 2004.

[43] JM Vilar, CC Guet, and S Leibler. Modeling network dynamics: the lac operon, a case study. *J Cell Biol.*, 161:471–6, 2003.

[44] M. Wagner, J. Meller, and R. Elber. Large-scale linearprogramming techniques for the design of protein folding potentials. Math. Program., in press.

[45] A. Warshel and M. Levitt. Theoretical studies of enzymatic reactions: dielectric electrostatic and steric stabilization of carbonium ion in the reaction of lysozyme. *J. Mol. Bio.*, 103:227–249, 1976.

[46] J. Watson et al. *Molecular Biology of the Gene.* Pearson, Bejamin Cummings, 2004.

[47] Y Wei, J-M Lee, C Richmond, FR Blattner, JA Rafalski, and RA LaRossa. High-density microarray-mediated gene expression profiling of e. coli. *J. Bacteriol.*, 183:545–56, 2001.

[48] A. J. Williamson, R. Q. Hood, and J. C. Grossman. Linear-scaling quantum monte carlo calculations. *Phys. Rev. Lett.*, 87:246406, 2001.

[49] J. Wooley. Frontiers at the interface between computing and biology. Technical report, NRC, Washington, DC, 2004.

[50] V. Zaloj and R. Elber. Parallel computations of moleculardynamics trajectories using stochastic path approach. *Comp. Phys. Commun.*, 128:118–127, 2000.

DISTRIBUTION LIST

Director of Space and SDI Programs
SAF/AQSC
1060 Air Force Pentagon
Washington, DC 20330-1060

CMDR & Program Executive Officer
U S Army/CSSD-ZA
Strategic Defense Command
PO Box 15280
Arlington, VA 22215-0150

DARPA Library
3701 North Fairfax Drive
Arlington, VA 22203-1714

Department of Homeland Security
Attn: Dr. Maureen McCarthy
Science and Technology Directorate
Washington, DC 20528

Assistant Secretary of the Navy
(Research, Development & Acquisition)
1000 Navy Pentagon
Washington, DC 20350-1000

Principal Deputy for Military Application [10]
Defense Programs, DP-12
National Nuclear Security Administration
U.S. Department of Energy
1000 Independence Avenue, SW
Washington, DC 20585

Superintendent
Code 1424
Attn: Documents Librarian
Naval Postgraduate School
Monterey, CA 93943

DTIC [2]
8725 John Jay Kingman Road
Suite 0944
Fort Belvoir, VA 22060-6218

Strategic Systems Program
Nebraska Avenue Complex
287 Somers Court
Suite 10041
Washington, DC 20393-5446

Headquarters Air Force XON
4A870 1480 Air Force Pentagon
Washington, DC 20330-1480

Defense Threat Reduction Agency [6]
Attn: Dr. Arthur T. Hopkins
8725 John J. Kingman Rd
Mail Stop 6201
Fort Belvoir, VA 22060-6201

IC JASON Program [2]
Chief Technical Officer, IC/ITIC
2P0104 NHB
Central Intelligence Agency
Washington, DC 20505-0001

JASON Library [5]
The MITRE Corporation
3550 General Atomics Court
Building 29
San Diego, California 92121-1122

U. S. Department of Energy
Chicago Operations Office Acquisition and
Assistance Group
9800 South Cass Avenue
Argonne, IL 60439

Dr. Jane Alexander
Homeland Security: Advanced Research
Projects Agency, Room 4318-23
7th & D Streets, SW
Washington, DC 20407

Dr. William O. Berry
Director, Basic Research ODUSD(ST/BR)
4015 Wilson Blvd
Suite 209
Arlington, VA 22203

Dr. Albert Brandenstein
Chief Scientist
Office of Nat'l Drug Control Policy Executive
Office of the President
Washington, DC 20500

Ambassador Linton F. Brooks
Under Secretary for Nuclear Security/
Administrator for Nuclear Security
1000 Independence Avenue, SW
NA-1, Room 7A-049
Washington, DC 20585

Dr. James F. Decker
Principal Deputy Director
Office of the Director, SC-1
Room 7B-084
U.S. Department of Energy
1000 Independence Avenue, SW
Washington, DC 20585

Dr. Patricia M. Dehmer [5]
Associate Director of Science for Basic Energy
Sciences, SC-10/Germantown Building
U.S. Department of Energy
1000 Independence Ave., SW
Washington, DC 20585-1290

Ms. Shirley A. Derflinger [5]
Technical Program Specialist
Office of Biological & Environmental Research
SC-70/Germantown Building
U.S. Department of Energy
1000 Independence Ave., SW
Washington, D.C. 20585-1290

Dr. Martin C. Faga
President and Chief Exec Officer
The MITRE Corporation
Mail Stop N640
7515 Colshire Drive
McLean, VA 22102

Mr. Dan Flynn [5]
Program Manager
DI/OTI/SAG
5S49 OHB
Washington, DC 20505

Ms. Nancy Forbes
Homeland Security Institute
Threats Division
2900 N. Quincy St. #600
Arlington, VA 22206

Dr. Paris Genalis
Deputy Director
OUSD(A&T)/S&TS/NW
The Pentagon, Room 3D1048
Washington, DC 20301

Mr. Bradley E. Gernand
Institute for Defense Analyses
Technical Information Services, Room 8701
4850 Mark Center Drive
Alexandria, VA 22311-1882

Dr. Lawrence K. Gershwin
NIC/NIO/S&T
2E42, OHB
Washington, DC 20505

Brigadier General Ronald Haeckel
U.S. Dept of Energy
National Nuclear Security Administration
1000 Independence Avenue, SW
NA-10 FORS Bldg
Washington, DC 20585

Dr. Theodore Hardebeck
STRATCOM/J5B
Offutt AFB, NE 68113

Dr. Robert G. Henderson
Director, JASON Program Office
The MITRE Corporation
7515 Colshire Drive
Mail Stop T130
McLean, VA 22102

Dr. Charles J. Holland
Deputy Under Secretary of Defense Science
& Technology
3040 Defense Pentagon
Washington, DC 20301-3040

Dr. Bobby R. Junker
Office of Naval Research
Code 31
800 North Quincy Street
Arlington, VA 22217-5660

Dr. Andrew F. Kirby
DO/IOC/FO
6Q32 NHB
Central Intelligence Agency
Washington, DC 20505-0001

Dr. Anne Matsuura
Army Research Office
4015 Wilson Blvd
Tower 3, Suite 216
Arlington, VA 22203-21939

Mr. Gordon Middleton
Deputy Director
National Security Space Architect
PO Box 222310
Chantilly, VA 20153-2310

Dr. Julian C. Nall
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

Dr. C. Edward Oliver [5]
Associate Director of Science for Advanced
Scientific Computing Research
SC-30/Germantown Building
U.S. Department of Energy
1000 Independence Avenue, SW
Washington, DC 20585-1290

Mr. Raymond L. Orbach
Director, Office of Science
U.S. Department of Energy
1000 Independence Avenue, SW
Route Symbol: SC-1
Washington, DC 20585

Mr. Thomas A. Pagan
Deputy Chief Scientist
U.S. Army Space & Missile Defense Command
PO Box 15280
Arlington, Virginia 22215-0280

Dr. Ari Patrinos [5]
Associate Director of Science for Biological
and Environmental Research
SC-70/Germantown Building
US Department of Energy
1000 Independence Avenue, SW
Washington, DC 20585-1290

Dr. John R. Phillips
Chief Scientist, DST/CS
2P0104 NHB
Central Intelligence Agency
Washington, DC 20505-0001

Records Resource
The MITRE Corporation
Mail Stop D460
202 Burlington Road, Rte 62
Bedford, MA 01730-1420

Dr. John Schuster
Submarine Warfare Division
Submarine, Security & Tech Head (N775)
2000 Navy Pentagon, Room 4D534
Washington, DC 20350-2000

Dr. Ronald M. Sega
DDR&E
3030 Defense Pentagon, Room 3E101
Washington, DC 20301-3030

Dr. Alan R. Shaffer
Office of the Defense Research and Engineering
Director, Plans and Program
3040 Defense Pentagon, Room 3D108
Washington, DC 20301-3040

Dr. Frank Spagnolo
Advanced Systems & Technology
National Reconnaissance Office
14675 Lee Road
Chantilly, Virginia 20151

Mr. Anthony J. Tether
DIRO/DARPA
3701 N. Fairfax Drive
Arlington, VA 22203-1714

Dr. Bruce J. West
FAPS - Senior Research Scientist
Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

Dr. Linda Zall
Central Intelligence Agency
DS&T/OTS
3Q14, NHB
Washington, DC 20505-00